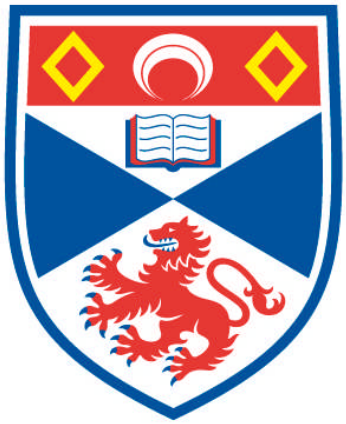
MACHINE LEARNING FOR SYSTEMS PATHOLOGY

(CHAPTERS 3 & 5 EMBARGOED)

Wim Verleyen

A Thesis Submitted for the Degree of PhD
at the
University of St Andrews



2013

University of
St Andrews

School of Biology

PhD Thesis

Machine Learning for Systems Pathology

by
Wim Verleyen

31st of July

Abstract

Systems pathology attempts to introduce more holistic approaches towards pathology and attempts to integrate clinicopathological information with "-omics" technology. This doctorate researches two examples of a systems approach for pathology: (1) *a personalized patient output prediction for ovarian cancer* and (2) *an analytical approach differentiates between individual and collective tumour invasion*.

During the personalized patient output prediction for ovarian cancer study, clinicopathological measurements and proteomic biomarkers are analysed with a set of newly engineered bioinformatic tools. These tools are based upon feature selection, survival analysis with Cox proportional hazards regression, and a novel Monte Carlo approach. Clinical and pathological data proves to have highly significant information content, as expected; however, molecular data has little information content alone, and is only significant when selected most-informative variables are placed in the context of the patient's clinical and pathological measures. Furthermore, classifiers based on support vector machines (SVMs) that predict one-year PFS and three-year OS with high accuracy, show how the addition of carefully selected molecular measures to clinical and pathological knowledge can enable personalized prognosis predictions. Finally, the high-performance of these classifiers are validated on an additional data set.

A second study, an analytical approach differentiates between individual and collective tumour invasion, analyses a set of morphological measures. These morphological measurements are collected with a newly developed process using automated imaging analysis for data collection in combination with a Bayesian network analysis to probabilistically connect morphological variables with tumour invasion modes. Between an individual and collective invasion mode, cell-cell contact is the most discriminating morphological feature. Smaller invading groups were typified by smoother cellular surfaces than those invading collectively in larger groups. Interestingly, elongation was evident in all invading cell groups and was not a specific feature of single cell invasion as a surrogate of epithelial-mesenchymal transition. In conclusion, the combination of automated imaging analysis and Bayesian network analysis provides an insight into morphological variables associated with transition of cancer cells between invasion modes. We show that only two morphologically distinct modes of invasion exist.

The two studies performed in this thesis illustrate the potential of a systems approach for pathology and illustrate the need of quantitative approaches in order to reveal the system behind pathology.

Declarations

I, Wim Verleyen, hereby certify that this thesis, which is approximately 72,500 words in length, has been written by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree. I was admitted as a research student in August 2008 and as a candidate for the degree of PhD in August 2009; the higher study for which this is a record was carried out in the University of St Andrews between 2009 and 2012.

date_____ signature of candidate

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

date_____signature of supervisor

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that my thesis will be electronically accessible for personal or research use unless exempt by award of an embargo as requested below, and that the library has the right to migrate my thesis into new electronic forms as required to ensure continued access to the thesis. We have obtained any third-party copyright permissions that may be required in order to allow such access and migration, or have requested the appropriate embargo below.

The following is an agreed request by candidate and supervisor regarding the electronic publication of this thesis: Embargo on both Chapter 3 and 5 of printed copy and electronic copy for the same fixed period of 2 years on the following ground(s): publication would preclude future publication.

date_____ signature of candidate

date_____signature of supervisor

Acknowledgements

During the course of my PhD studies, there were many people that had a great contribution to the outcome of my work.

First, I would like to thank my main supervisor V. Anne Smith. She has been a great mentor during my time in St Andrews. I am very grateful for many opportunities to follow my scientific interests. I really enjoyed the confidence and support in situations I was not so sure about the outcomes of my work. I really need to thank her for her persistent patience and inspiring advice that have defined a standard for me which I will try to emulate when I supervise students.

To my thesis committee, thank you for being engaged during the process of my doctorate. I have been fortunate to receive guidance, and plenty of suggestions which have inspired me to improve my research into many directions.

I need to thank the Division of Pathology at the University of Edinburgh for their contribution. Prof. David Harrison, Dr. Dana Faratian, and many more have been treating me as being a part of their research group whenever I needed their advice, infrastructure, and data.

I would like to thank SIMBIOS at the University of Abertay Dundee. Prof. James Bown, Dr. Mark Showman, Mr. Michael Idowu, and many more to be always very knowledgeable and informative in the course of this study.

I would like to convey thanks to the Scottish Universities Life Sciences Alliance (SULSA), the School of Biology of the University of St Andrews, and Prof. David Harrison for funding my doctorate studies.

Finally, I want to thank my family and friends for their love and support. And most of all for my supportive, encouraging, and inspiring girlfriend Olivia for her patient and understanding during the final stages of the Ph.D. is so appreciated.

Wim Verleyen
University of St Andrews
August 2012

Contents

List of Figures

List of Tables

1
Introduction

> By the deficits we may know the talents, by the exceptions we may know the rules, by studying pathology we may construct a model of health.
>
> *Laurence Miller*

Cancer is a collection of different diseases. This collection of diseases results in the heterogeneity of cancer, dynamic biological processes, and adaptive response to therapy.

Pathology allows us to classify cancer in different categories based on stage, morphology, histology, etc. These pathological characteristics have shown to help the diagnosis of cancer in the clinic.

Generally, the prediction of therapeutic outcome of a patient with cancer is still very challenging and in need of new approaches for its inference. Systems pathology introduces a holistic approach towards building the model of health of a patient. This model is not only based on the more traditional pathological measures (i.e., stage or histological state), but it can be extended with "omics" technology available for generating data of genomes, exomes, proteomes, transcriptomes. metabolomes, etc.

In the following sections of this chapter, machine learning will be introduced. This will be followed by an description of the term systems pathology. It will be followed by a description of cancer, and goes into more detail on breast- and ovarian cancer. This introduction chapter reflects the context of this PhD, it studies machine learning algorithms to build models for the pathology of cancer.

I have constructed a novel set of tools for computational biology and bioinformatics based on machine learning algorithms for the integration of complex heterogeneous data. This novel tools are engineered based on high-throughput data collected by my collaborators and myself. This unique systems approach extends more traditional approaches for answering fundamental questions in pathology.

1.1 *Machine learning*

Machine learning provides an independent *"in silico"* representation of learning based on experimental data [Solomonoff, 1956]. This learning involves recognizing patterns in the underlying distribution of this data [Bishop, 2006a].

Machine learning can be defined as a computer program that learns from experience (i.e. data), with respect to a task (i.e. classification of patients as having a high risk of recurrence of cancer within one year, or death within three years, etc.) and some performance measure (i.e. error rate, accuracy, precision, etc.); its performance on the task, as measured by the performance measure, improves with experience [Mitchell, 1997].

Machine learning, and also statistical modelling can be categorized into: *generative*, and *discriminative* machine learning [Bishop, 2006b, Jebara, 2002]. Generative machine learning models the probability density over all variables (joint probability distribution), e.g. mixture models, Markov logic networks, hidden Markov models, Bayesian networks etc. Discriminative machine learning specifies strategies to model direct mappings between input, and output variables (conditional probability distributions), e.g. logistic regression, Gaussian processes, support vector machines, etc.

Furthermore, machine learning is divided into three types of learning [Bishop, 2006c]:

1. *supervised learning*: given a set of input vectors, training data, and output vectors, learn a function ($f(x)$) between input and output vectors. These learners are used for classification (discrete output vector) and regression (continuous output vector).

2. *unsupervised learning*: the data only contains attributes, input- and output vectors are not distinguished. Examples of these learners are clustering, density estimation, and visualization.

3. *reinforcement learning*: a specific type of learner that involves taking actions depending on the input vectors and its environment. Specific actions are rewarded and punished depending on how they are quantified in the learner.

I would like to make a remark upon these categories and types of machine learning. Many machine learning algorithms combine these types and categories, e.g. Bayesian networks are a generative machine learning approach, and are called unsupervised learning when we perform structure learning.

In the following sections, probabilistic models and survival analysis will be introduced. More theoretical background and applications of various machine learning algorithms can be found in chapters 2, 3, 4, and 5.

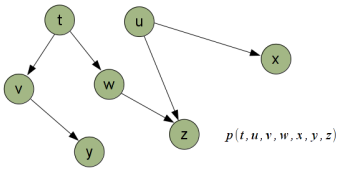


Figure 1.1: Schematic view of generative machine learning.

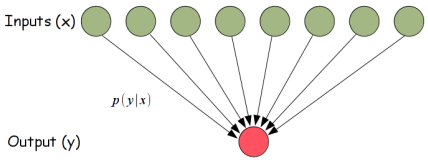


Figure 1.2: Schematic view of discriminative machine learning.

1.1.1 Probabilistic graphical models

Probabilistic graphical models combine *uncertainty* (probability theory) and *graphical structure* (independence constraints). It is a very general approach to construct statistical models (Kalman filters, hidden Markov models, Isling models, etc.) into a graphical representation.

There are three main types of probabilistic graphical models: (1) *Bayesian networks* (also called *belief networks* or *causal networks*) [Pearl, 1988], (2) *mutual information networks* [Meyer et al., 2008], and (3) *Markov networks* (also called *Markov random fields (MRFs)*) [Getoor and Taskar, 2007]. Bayesian networks are *directed graphical models*, and mutual information networks and Markov networks are *undirected graphical networks*.

Probabilistic graphical models provide a picture of the joint probability distribution over a set of random variables ($\chi = \{X_1, X_2, \ldots, X_n\}$). The structure is a representation of the independence properties of our system under investigation. These independence properties represent a high-dimensional joint probability into a compact and coherent manner.

Probabilistic graphical models are part of artificial intelligence (AI). AI research is concentrated in two major disciplines: (1) *logical representation* (logic programming, description logic, classical planning, symbolic parsing, rule induction, etc.) and (2) *statistical - uncertainty representation* (Bayesian networks, hidden Markov models, Markov decision processes, statistical parsing, neural networks, etc.). These two major disciplines of AI are combined into one framework, called *Markov logic* [Richardson and Domingos, 2006].

1.1.2 Survival analysis

Survival analysis is applied to describe and quantify time-to-event data [Stevenson, 2009]. This group of approaches focuses on the distribution of survival time (T). It has been successfully used for different types of problems: time-to-death analysis, time-to-event analysis in sociology, etc. Data collections in a biomedical environment often contains with a follow-up time dimension. The start point and end point of this follow-up period could lead to incomplete information. This is called *censoring*. There are three mean types of censoring: (1) *right censoring*: if the event of interest occurs *after* the recorded follow-up period, e.g. a patient is still alive after the period of observation, (2) *left censoring*: if the event of interest occurs *before* the recorded follow-up period, e.g. when the initial risk is unknown, and (3) *interval censoring*: when left and right censoring occur together.

There are different statistical approaches to perform survival analysis, e.g. non-parametric, parametric, and semi-parametric. The non-parametric approaches are often used as a starting point, they can be used to plot the survival time distribution, and allow us to make comparisons between different categories in the data

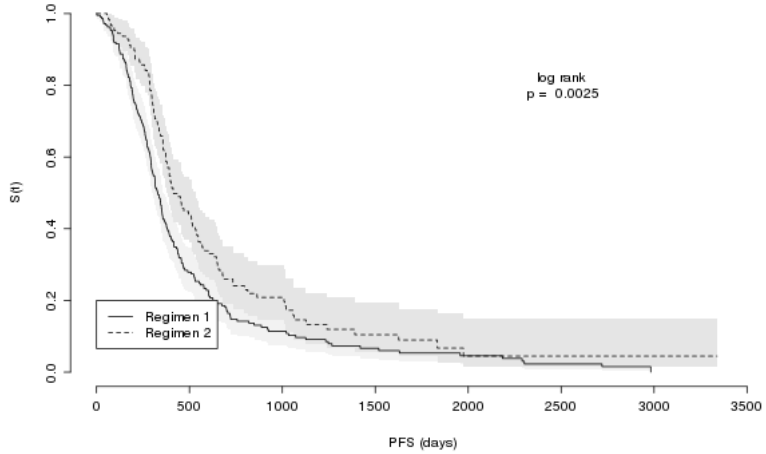


Figure 1.3: An example of the survival function for the progression-free survival (PFS) for patients under different treatment regimens (Regimen 1: platinum and Regimen 2: platinum combined with taxane). The grey area indicates the 95% confidence interval (see also section 2.1.2.1 on page 49).

set (see figure 1.3). Examples of non-parametric survival analysis are Kaplan-Meier, life table method, Nelson-Aalen method, and Flemington-Harrington method [Collett, 2004a]. Based on this survival time distribution a parametric approach can be applied. Finally, the approaches that we will concentrate on during this research are the semi-parametric approaches, e.g. Cox proportional hazards regression, partial Cox regression, and survival support vector machines.

1.2 Systems pathology

The twentieth century is often called the *century of physics*; the twenty-first century is often predicted as the *century of biology* [Sengupta, 2006]. The current trend in biology is to adapt engineering concepts to have a holistic approach to "what is biology"?, moreover, to "what is cancer"? The research completed during this thesis deals with the computational side of biology. It is a crosspoint of mathematics, computer science and biology. Research at this intense combination of fields can be named *systems biology*. It can also be named *computational biology* or *bioinformatics*.

Systems biology is not completely new. Its origin lies in the end of the '50s, and the beginning of the '60s. A pioneer in systems biology is Denis Noble [Noble, 2006], who is a British biologist - physiologist. He introduced the first computer model of a *virtual heart* during his PhD [Noble, 1961] at *University College London* in 1961. Because of gaps in knowledge and the very complex dynamics of biological systems, systems biologists are still seeking new methodologies, mathematical modelling, and software. These new findings can have a major influence on how biologists have new insights in their own field [Westerhoff and Palsson, 2004, Palsson, 2006].

Biology nowadays is often modelled by pathways. Different interactions between pathways are often called *networks*. In essence all relationships can be modelled as a graph. As Bayesian networks use graphs, Bayesian networks are a good tool for modelling networks. Networks are a very popular representation of the system under investigation. Albert-László Barabási [Barabási, 2003] had a major influence in pointing out general concepts in networks. He is the inventor of the *scale-free networks* [Barabási and Albert, 1999]. They can be found in many fields, and also play a fundamental role in biology, sociology, computer science, simulation, economics, etc. Biology can be represented as networks of biological interactions [Barabási and Oltai, 2004]. A genetic mutation can lead to a modification of these interactions and can be a cause of cancer. Cancer drugs can target specific nodes in biological network as a strategy for drug development [Barabási, 2003]. A special antibody is provided to a tumour cell to interact with the cancer biology networks to get cancer cells into *apoptosis* and the tumour cell disappears [Azim and Jr., 2008].

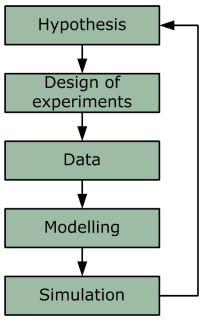


Figure 1.4: Systems biology block scheme.

This new systems biology trend has gained an ever increasing interest during the last decade. Often systems biology is a loop of five successive tasks [Palsson, 2006]: (1) hypothesis, (2) design of experiment, (3) data, (4) modelling, and (5) simulation. After the simulation results are analysed, the hypothesis can be tested, and eventually can lead to a new hypothesis (see figure 1.4). Nevertheless, there are still prominent biologists that are very sceptical of the feasibility of systems biology approaches, e.g. the famous quote by Sydney Brenner: "low input, high throughput, no output" [Brenner, 2010]. Systems biology attempts to test more holistic hypotheses, which implies an increased amount of complexity that needs to be

explained. A systems approach can certainly be a step forward in understanding biology, but can also lead to wrong conclusions, i.e. if one misformulates a system specific question, a mathematical model constructed from a large data set might not provide very meaningful output.

In medicine a disease is often defined by its *etiology* (cause), its clinical observable signs and symptoms, its *pathogenesis* (underlying mechanisms that cause the signs and symptoms), its natural history, and its treatment [Hunter, 2009]. Pathology covers all this information, and is the study and the diagnosis of a disease.

Since molecular biology and pathology can be studied with high throughput "omics" technology this research field goes through the transition from qualitative to quantitative. This transition can be explained by the change in data resources (i.e., many pathological features are qualitative and new imaging analysis can provide quantitative measures). These novel data collections allow to answer more holistic hypotheses and is, by analogy with systems biology, called systems pathology [Faratian et al., 2009].

1.2.1 Clinicopathological definition of a tumour

The histopathological definition of tumour tissue has important implications on cancer progression and treatment. Microscopic examination remains the primary diagnostic method for tumour tissue [Cesario and Marcus, 2011]. The nomenclature of tumours are based either on *histogenesis* or *histology*. The histogenesis studies the tissue of origin during development and formation of a tumour. Histology describes anatomic properties of the tumour tissue. In cancer, histology often compares the tissue under diagnosis with normal tissue.

Tumour tissues are constructed of two parts: (1) *tissue neoformation (parenchyma)* and (2) *stroma* [Kalluri and Weinberg, 2009, Hong et al., 2010]. The neoformed tissue appears in two forms: (1) *carcinoma*, epithelial cells that form internal and external body surfaces and cavities, and (2) *sarcoma*, mesenchymal cells that form more connective tissue, e.g. bone, lymphatic, cartilage, etc. Stroma is important for the different biological programs active in a tumour [Beck et al., 2011]. It is a reservoir for the tumour to find new cells or an environment that provides resources for tumour invasion [Kalluri and Zeisberg, 2006].

1.2.1.1 Tumour grading

A tumour appears in two main types: (1) *benign* or (2) *malignant*. Sometimes a tumour is defined into intermediate state, i.e. semi-malignant, pseudo-malignant, and of questionable malignancy [Cesario and Marcus, 2011]. Benign neoplasms[1] grow slowly, do not harm[2], and are non-invasive. Whereas, malignant tumours are characterized by high proliferation, invade tissue, and metastasize [Ludwig and Weinstein, 2005].

[1] Neoplasm is a more accurate nomenclature as tumour. A neoplasm is an abnormal construction of tissue resulting from neoplasia; neoplasia is the proliferation of cells. A tumour is a less specific definition of a swelling.

[2] Benign tumours can do harm, e.g. craniopharyngiomas or pituitary tumours can press on the optic chiasm impairing vision or causing blindness.

The grade of a tumour defines the activity of a malignant tumour. This activity specifies rate of growth, stromal reaction, and differentiation of cells as a measure of cancer progression.

Table 1.1 lists a brief overview of the tumour nomenclature[3] based on cell origin, mixed tumour tissues, and cell secretory activity[4].

[3] Most benign tumours are have the suffix -oma; there are exceptions, e.g. melanoma, seminoma, etc.

[4] Cell secretory activity results in emition of chemicals of a cell.

Table 1.1: Histogenetic nomenclature of tumour tissue [Hamilton and Aaltonen, 2000, Chan, 2001]

Description	Benign tumour nomenclature
Epithelial tumour originates from glandular tissue (i.e., gastro-intestinal tract, breast. kidney, liver, etc.)	*Adenoma*
Non-secretory epithelial surfaces (i.e., skin, respiratory mucosa, lower urinary tract, etc.)	*Papilloma*
Mesenchymal tumour with fibroblasts	*Fibroma*
Mesenchymal tumour with adipocytes	*Lipoma*
Mesenchymal tumour with osteoblasts	*Osteooma*
Mixed epithelial-mesenchymal tumour	*Fibropapilloma, adenofibroma*
Cell secretory activity	*Mucinous, colloid, serous, apocrine, or neuroendocrine*

Tumour grading and its histological typing varies among different types of cancer [Hong et al., 2010]. Histopathological types have a standard nomenclature [Hamilton and Aaltonen, 2000, Chan, 2001]. The definition of grade is complicated by the heterogeneity of certain neoplasms; its mapping between a histological type and clinical observation can be partial.

1.2.1.2 Tumour staging

Tumour staging is an important measure for deciding on treatment. Tumour staging classifies a tumour based upon the spread and size of the neoplasm [Hong et al., 2010].

An international used staging is called *Tumour Node Metastasis (TNM) staging system*. This system has three parameters: (1) T, represents the size of the primary tumour and its behaviour towards surrounding structures, e.g. adjacent, in contact, or invasive, (2) N indicates involvement not important of regional lymph nodes, and (3) M specifies if metastasis exists. There are two main versions of the TNM staging system. One is designed by the International Union Against Cancer (UICC) [Greene and Sobin, 2009], and the other by the American Joint Comittee on Cancer (AJCC) [Edge and Compton, 2010].

Tumour TNM classification also specifies four stages [Greene and Sobin, 2009]: (1) *stage 1*, tumour invades muscularis propria, but has not spread to nearly lymph nodes, (2) *stage 2*, tumour spreads into the subserosa and/or perirectal tissues with up to three regional lymph nodes, or directly invades adjacent tissues without lymph node involvement, (3) *stage 3*, any depth of tumour invasion, with four or more positive lymph nodes, but without distant metastasis, (4) *stage 4*, any depth of tumour involvement, any number of involved lymph nodes, with distant metastasis.

Tumour staging is the most powerful, and well standardized, diagnostic measure in the clinical environment [Cesario and Marcus, 2011]. Despite its success, it does not capture detailed morphological characteristics of a neoplasm, information related to the dynamics of tumour invasion, etc. Therefore, more holistic systems approaches are required to discover more complete clinical measurements for better prognosis. There is huge potential for combining computational modelling with biomarkers as a first step to progress to a more personalized therapy.

1.3 Cancer

Cancer is heterogeneous disease, and characterized by fundamental biological processes, e.g. *cell regulation*, *cell proliferation*[5]: *cell growth* and *cell division*, *cell differentiation*, etc. [Hong et al., 2010]. Extra-and intracellular communication in normal cells leads to *homeostatic* mechanisms. Cancer cells, depending on the stage of tumour formation, are in a certain degree of disequilibrium. From an evolutionary point of view, biological systems are fairly robust[6] [Wagner, 2010]. This *robustness* is tested by the occurrence of specific phenotypes. Neoplastic diseases can be driven by *proliferation*. This proliferation is often, maybe even always, characterized by a disordered cell differentiation; in tumourigenesis it leads to the construction of *anaplasia*[7] [Hong et al., 2010].

[5] cell proliferation results in an increased number of cells, and is often a combination of cell growth, and cell division.

[6] The most robust biological systems have a higher probability to survive compared to less robust biological systems (i.e., survival of the fittest).

[7] Anaplasia are malignant neoplasms.

Figure 1.5 on page 21 pictures an overview of many fundamental biological interactions in cancer. Depending on the type, and subtypes of cancer, etc. different interactions have more importance for diagnosing and treating cancer in a more sophisticated approach. The following sections provide an overview of the most important known mechanisms in cancerous cells. Some of these mechanisms are explained in a frequently cited work: *The hallmarks of cancer*, it has been recently revised [Hanahan and Weinberg, 2000, Hanahan and Weinberg, 2011]. A hallmark can be defined as a feature of a system that differentiates it from other systems [Yuri, 2010]. These features will be introduced together with the corresponding biological implications. As a general guidance table 1.2 on page 22 provides a summary of the six hallmarks.

These hallmarks are often a consequence of mutations in a cell. These mutations lead to two main types of oncogenes: (1) *proto-oncogene* and (2) *suppressor oncogene*. These oncogenes can be found by a microarray experiment; the oncogenes will differentially expressed for different biological conditions, e.g. compare the gene expression levels of benign tumours with malignant tumours. A question that arises when measuring these expression levels for the discovery of these hallmarks is: "What is the negative control for defining that an expression level is qualitatively low or high?". In the previous microarray experiment, the negative control could be a *benign* tumour. So, expression levels of malignant tumours can be compared against benign tumours [Yuri, 2010].

The following paragraphs will introduce the global statistics of cancer and fundamental cancer terminology. Some of these terms will reoccur in later chapters, other terms are fundamental for understanding the literature.

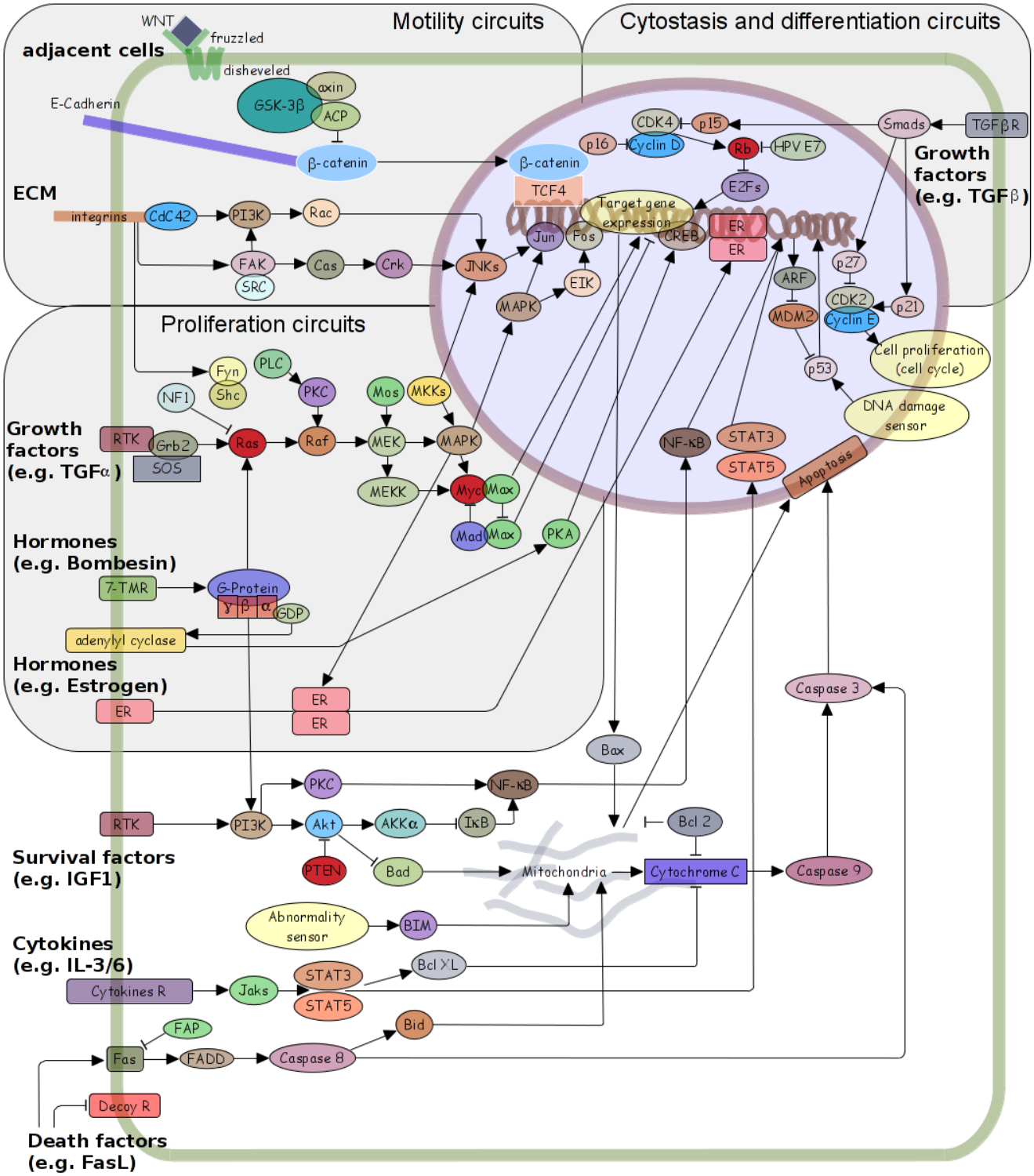


Figure 1.5: An overview of of known biological circuits active during cancer. The figure is constructed from [Hanahan and Weinberg, 2000, Hanahan and Weinberg, 2011].

Table 1.2: Overview of the six hallmarks of cancer. The table is constructed from [Hanahan and Weinberg, 2000, Hanahan and Weinberg, 2011].

Hallmark	Pathways	Interpretation
1. Sustaining proliferative signaling	growth factor ligand Ras signal transducer	normal tissue poorly understood
somatic mutations activate downstream pathways	B-Raf, Raf and, MAPK signaling	crosstalk among different pathways
	PI3K and Akt/PKB signaling	
disruptions of negative feedback loop	Ras mutation induces Ras GTPase activity	homeostatic regulation, and drug resistance development
	PTEN phosphatase, PI3K, and PIP_3	loss of function mutations in PTEN amplify PI3K
	mTOR in up-and downstream of PI3K, and Akt/PKB	anti-proliferative effects of mTOR inhibition
excessive proliferation can trigger cell senescence	Ras, Myc, and Raf	too high (Ras) expression can invoke cell senescence and apoptosis, too low invokes proliferation
2. Evading growth suppressors	Rb, and P53 signaling	key targets to make the cell proliferate, or induce cell senescence and apoptosis very complicated wiring to activate this fundamental targets
mechanisms of contact inhibition and its evasion	NF2, Merlin, E-Cadherin, and RTK	homeostatic regulation
	LKB1 and Myc	LKB1 functions as a suppressor for excessive proliferation many more to be discovered
corruption TGFβ pathway promotes malignancy	TGFβ pathway signaling	anti-proliferate effects in late-stage tumours it activates epithelial-to-mesenchymal transition (EMT)
3. Resisting cell death		extra-and intracellular apoptosis inducing circuits
	Bcl-2 family of regulatory proteins: Bcl-X_L, Bcl-w, Mcl-1, and A1	inhibitors of apoptosis
	Bak and Bax (share BH3 domains)	proapoptotic proteins; BH3 domain induces Bcl-2 apoptosis inhibition or apoptosis
	DNA damage sensor that functions via P53	P53 induces apoptosis by up-regulation of Noxa and PUMA BH3 only proteins; this is a response DNA breakage and chromosomal abnormalities
	Bim BH3 only domain protein	survival factor signaling
	Myc	target for apoptosis via Bim and other BH3-only domain proteins
autophagy mediates tumour cell survival and death	PI3K, Akt, and mTOR signaling	survival factor signaling blocks apoptosis and autophagy research needed to discover physiological/genetic features that cause autophagy to die or survive cancer cells
	Beclin-1 BH3 only domain protein	targets autophagy, and can induce apoptosis via Bax/Bak stress transducing BH3 proteins (Bid, Bad, Puma, etc.) potentially induce apoptosis/autophagy
necrosis: pro-inflammatory/tumour-promoting potential	IL-1α	stimulate tumour growth potential
4. Enabling replicative immortality	telomeres are protecting the end of chromosomes are centrally involved in the capacity of unlimited proliferation	cells continue to proliferate after senescence and bypass crisis into immortalization
reassessing replicative senescence		cell senescence remains a barrier for proliferation; it can depend on the cell culture conditions.
delayed activation of telomerase may both limit and foster neoplastic progression	telomerase and P53	lack of telomerase and P53 function
new functions of telomerase	TERT, Wnt, and βCatenin/LEF	additional functions need to be discovered
5. Inducing angiogenesis	VEGF signaling, MMP-9, FGF, TSP-1	important in early development stages of tumours
	TSP-1	key target in the angiogenic switch
key gradations in the angiogenic switch	VEGF, Ras, and Myc	induction of angiogenesis can also stimulation proliferation
endogenous angiogenesis inhibitors present natural barriers to tumour angiogenesis	TSP-1, fragments of plasmin angiostatin and type 18 colagen endostatin	endogenous inhibitors of angiogenesis
pericytes are important components of the tumour neovasculature		
variety of bone marrow-derived cells contribute to tumour angiogenesis		
6. Activating invasion and metastasis	E-Cadherin	loss of E-Cadherin in carcinoma cells; key cell-to-cell adhesion molecule
EMT program broadly regulates invasion and metastasis	transcription factors Snail, Slug, Twist, and Zeb1/2 organize EMT function	EMT supports invasion, avoids apoptosis, and disseminate
heterotypic contributions of stromal cells to invasion and metastasis	IL-4, EGF, and CSF-1	crosstalk between cancer cells and cells of neoplastic stroma induces invasiveness and metastatis
plasticity in invasive growth program		Reverse of EMT: mesenchymal-epithelial transition (MET)
distinct forms of invasion may underlie different cancer types		characterization of different types of tumour invasion
daunting complexity of metastatic colonization		research need to reveal regulatory programs that define metastatic colonization

Global statistics In global death statistics in the US [Murphy et al., 2012], EU [EUD, 2012], and UK [UKCR, 2012] *diseases of heart (ICD: I00-I09,I11,I13,I20-I51)*[8], and *malignant neoplasms (ICD: C00-C97)* are the two main causes of death. Generally, diseases of the heart are the main cause of death, whereas, e.g. in the UK people aged above 50 [UKCR, 2012], and in US for people aged between 45 and 64, more than 30 % died of cancer [Murphy et al., 2012].

[8] ICD code is the International Classification of Disease coding system set by the World Health Organisation (WHO).

Proliferation In many cancers proliferation is a driving force [Hall and Levison, 1990, Schlabach et al., 2008]. Proliferation leads to an increase of the number of cells, and therefore closely related to cell growth and division. Since the cell is a robust system, there are mechanisms that limit this proliferation [Albert et al., 2002, Hong et al., 2010]. One of this mechanisms is called *senescence* (i.e. aging of a biological organism), another mechanism is called *apoptosis* (i.e cell death).

Differentiation Cell differentiation occurs when a cell in a multicellular organism is dedicated to become a specific cell type. The *memory of the cell* has a predefined biological genetic footprint for each cell [Albert et al., 2002]. In tumour cells this footprint is disordered; the same genes are part of this footprint, but their expression levels differ from the prototype cell [Hong et al., 2010].

Metastasis A general capacity of malignant tumours is to spread into different organs, this capacity is called *metastasis* [Albert et al., 2002, Fidler, 2003, Hong et al., 2010]. Tumours that are formed in the original organ are called *primary tumours*. Analogously, tumours that are spread into a different organ are called *secondary tumours*. The transition of primary towards secondary tumour is often called *colonization*. E.g., a bone metastasis drug called *Denosumab (Prolia®by Amgen. Inc.)*, has been approved by the FDA for cancer patients [Sethi and Kang, 2011].

Epithelial-mesenchymal transition (EMT) An important and not very well understood program during development of a biological system is called *epithelial-mesenchymal transition (EMT)* [Kalluri and Weinberg, 2009, Hong et al., 2010]. This program appears, with similar biological phenotypes, in three different classes of biological development: (1) EMT during implantation, embryogenesis, and organ development, (2) EMT associated with tissue regeneration and organ fibrosis, and (3) EMT associated with cancer progression and metastasis. In primary tumours EMT is active, whereas during colonization, the inverse process *mesenchymal-epithelial transition (MET)* is active [Hong et al., 2010]. The execution of EMT needs further investigation in order to improve insights in many biological processes, e.g. role of EMT during resistance to therapy, role during tumour invasion, etc. [Sanchez-García, 2009, Davidson et al., 2012].

Autophagy – angiogenesis Because of the proliferation and the spread of a tumour, a lot of energy is required to form a malignant tumour. There are two main energy related concepts active in tumour development: (1) *autophagy* (i.e. self-eating) [Kondo et al., 2005] and (2) *angiogenesis* (i.e. formation of new blood vessels) [Kalluri, 2003]. Both of these concepts can be seen as the required energy to run the switch of the transition between benign towards malignant tumours.

Autophagy can suppress tumour growth, but at the same time it can stimulate tumour growth by providing new energy [Hippert et al., 2006]. The energy often comes from the inner cells of a tumour [Mathew et al., 2007]. Despite this dual role of autophagy, it also plays a role in drug resistance, e.g. in HER2+ breast cancer treated with Herceptin (also called trastuzumab) autophagy markers were highly expressed [Vazquez-Martin et al., 2009].

() Angiogenesis, a term for the construction, repair, and migration of blood vessels, is another fundamental biological program in cancer [Folkman, 1995, Kullari, 2003]. This program has been mainly seen as a switch for the transition from primary to secondary tumours. This switch runs when the angiogenic phenotypes are triggered. Since the mid 1990s there is growing evidence that angiogenic tumour activity is fundamental for tumour growth and metastasis [Carmeliet, 2005, Sethi and Kang, 2011]. Recently, there has been multiple angiogenic inhibitors for the clinical practice, where side effects seem to be one of the biggest drawbacks. Since angiogenesis is a fundamental process in wound healing, heart function, reproduction of vessels, etc. If a drug inhibits this process, it is shown to induce toxicity [Verheul and Pinedo, 2007]. In the clinical environment the progression-free survival was extended, as there was little improvement of the overall survival figures (e.g., pazopanib in renal cell cancer). More holistic drug that integrates multiple strategies into an agent might introduce less toxicity in the future.

Cell growth and division cycle An important biological process in cancer is the cell growth and division cycle [Hartwell and Kastan, 1994, Clyde et al., 2006]. The cell cycle is an ordered sequence of events whereby a cell grows and then divides resulting in the production of two daughter cells that are identical to the original parent cell.

The cell cycle may be considered in five separate phases [Albert et al., 2002]:

1. **Gap one or G1 phase**: the cell undergoes a series of biochemical and physiological changes including sustained growth.

2. **Gap zero or G0 phase**: the cell is in a quiescent state and can be seen as a resting phase; often entered from a cell cycle checkpoint in the **G1 phase**. This phase occurs by a lack of mitogenic signal to proliferate. Most of the cells in a human body are in **G0** state.

3. **Synthetic or S phase**: the cell copies its DNA resulting in the development of duplicate copies of each chromosome.

4. **Gap two or G2 phase**: a second gap phase during which the proteins and complexes necessary for the remainder of the cell are synthesized.

5. **Mitosis or M phase**: *mitosis*, during which the cell divides with one set of chromosomes being allocated to each of the two resulting daughter cells.

Cell regulation Cell regulation is essential for preserving the appropriate functionality of living cells and maintaining a healthy phenotype [Clyde, 2006]. Regulation is maintained through a variety of gene expression processes which result in the supply of the proteins necessary for cell regulation at the correct time and in the correct quantities.

Cell regulation consists of a number of separate processes:

1. **Growth and division cycle**: this biological process occurs sequentially in separate steps leading to a terminally differentiated adult cell.

2. **Apoptopic pathways**: this intrinsic and extrinsic process regulates cellular death at the appropriate time and to the appropriate extent.

3. **Cell survival pathway and the anti-growth pathway**: this process is important in maintaining the balance of cells essential for homeostasis in multi-cellular species. A collection of cellular events, with an origin from different pathways, occur as a network.

4. **Damage response pathways**: this process monitors DNA damage and tries to repair DNA damage. If damage can not be repaired, cell will go to apoptosis (see category 2).

All these pathways link the cell's internal processes. Specific parts of these pathways are also linked to external intervention processes. This can be illustrated by the action of growth factor, and other ligands which can direct appropriate cell regulation as well as other forms of interaction with adjacent cells, or the extra-cellular matrix can interact in the cell regulation process.

In the following sections, we describe important facts related breast- and ovarian cancer.

1.3.1 Breast cancer

During 2008, almost a quarter of the female cancer deaths were due to breast cancer [Jemal et al., 2011]. Since the 1980s and 1990s, breast cancer has been slightly decreasing. This decrease is due to better early detection strategies and less postmenopausal hormone therapy [Jemal et al., 2011].

Breast cancer is usually not difficult to diagnose. The problem lies with stratification of patients for the right treatment, because of intrinsic or acquired resistance to therapy is hard to predict. A more successful therapy for a specific patient is still an enormous scientific challenge because of the *heterogenecity* of cancer [Sorlie et al., 2001, Med, 2008].

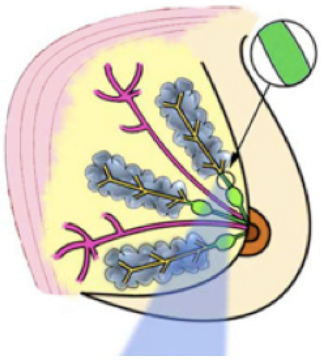


Figure 1.6: Mammary ductal network (Modified from: [Med, 2008]). This ductal network is analysed to detect the stage of breast cancer.

Classification of breast cancer Breast cancer is one of the best understood cancers [Gray and Druker, 2012]. Breast carcinomas are not only classified upon their histopathological measurements, Sorlie et al [Sorlie et al., 2001] (see figure 1.7 on page 27) illustrate, by applying hierarchical clustering, that breast cancer can be classified according to the gene expressions of cDNA microarray experiments into three major groups:

1. ER+ express typical protein of luminal epithelial cells.
   - luminal subtype A (most frequently occurring breast cancer).
   - luminal subtype B (second most frequently occurring breast cancer).
   - luminal subtype C (least frequently occurring breast cancer).
2. ER-
   - HER2+: ERBB2+.
     This tumour occurs in approximately 20 % of all breast cancers. HER2+ tumours tend to be more aggressive than HER2- tumours [Azim and Jr., 2008]. Furthermore, these tumours are characterized with ErbB2 gene amplification and HER2 receptor overexpresssion [Yarden and Sliwkowski, 2001, Hynes and MacDonald, 2009].
   - basal-like (15 % of all tumour carcinoma)
   - normal breast-like:
     This type of carcinoma has been regarded as a different type [Sorlie et al., 2006]. Its analysis it very complicated since it is similar to epithelial cells; its histological characteristics and clinical prognosis are still under research.

More recently, plenty of research is performed for the discovery of novel biomarkers and their acceptance into the clinical environment [McCafferty et al., 2009]. There are four main types of breast cancer: *luminal subtype A*, *luminal subtype B*, *basal-like*, and *HER2*; these types are characterized with four biomarkers: oestrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), and Ki67. The most frequently occurring breast cancers are *luminal subtype A*; they occur with following biological footprint: ER and PR positive, HER2 negative, and Ki67 low. Second most frequently occurring beast cancers are *luminal subtype*

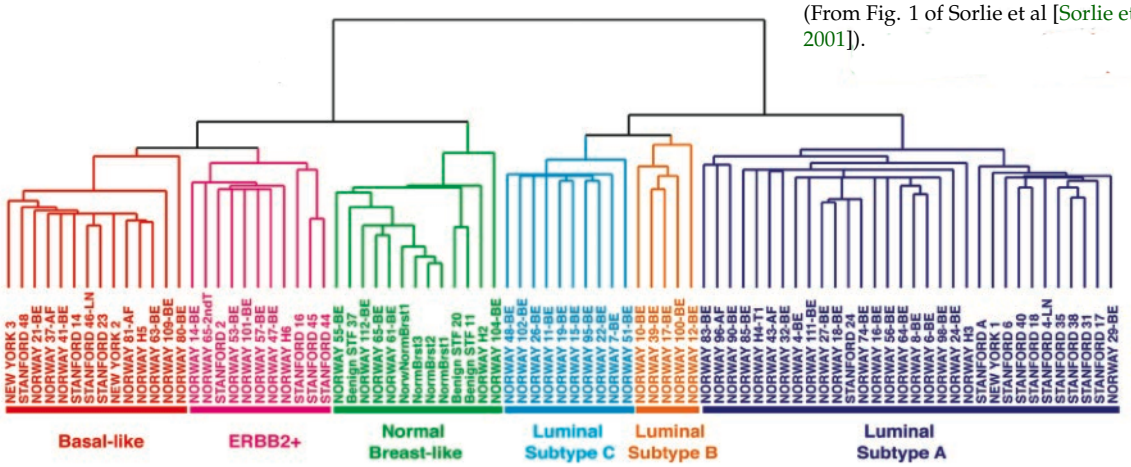


Figure 1.7: Hierarchical clustering applied for the gene expression data (From Fig. 1 of Sorlie et al [Sorlie et al., 2001]).

B; they occur with ER positive, PR and HER2 negative, and Ki67 high. HER2 breast cancer have HER2 positive, ER and PR negative, and Ki67 high. It becomes more complicated for the triple negative breast cancers (ER, PR, and HER2 negative and Ki67 high): basal-like.

For basal-like breast cancers are further subclassified [Rakha et al., 2008b]. Basal-like cancers can be identified with a positive expression of epidermal growth factor receptor (EGFR) and cytokeratin 5/6 (CK 5/6) biomarkers [Cheang et al., 2006]. This classification is not perfect, since triple negative basal-like tumours appear with negative expressions of EGFR and CK 5/6 and not all basal-like tumour appear with triple-negative signature [Rakha et al., 2008a].

Finally, *aprocrine type* are ER and PR negative and androgen receptor (AR) is positively expressed [Celis et al., 2009]. It is not clear if this is a class of breast tumours is distinct because of clinicopathological observations or can by part of any of the above described tumour classes [Gonzalez et al., 2008].

A systems approach could be beneficial to decipher the complex cascade of events that could further improve the treatment and management of breast cancer in the clinic [Barabási and Oltai, 2004, Nevins, 2007].

1.3.2 *Ovarian cancer*

Ovarian cancer is the fourth most common cancer death in the UK [UKCR, 2012], and the seventh most common cancer death in the US during 2008 [Jemal et al., 2011]. There are two main types of ovarian cancer: (1) *epithelial ovarian cancer (EOC)* and (2) *ovarian germ cell tumour*.

Epithelial ovarian cancer in early stage has a 5 year survival of ~90 %, but for late stage it is ~30 % [Lu et al., 2004, Faratian et al., 2011]. Currently there are no, or very limited, markers available for

early detection of epithelial ovarian cancer [Lu et al., 2004, Tothill et al., 2008]. Despite a need for markers of early detection, a set of markers is needed to understand the underlying biological mechanisms and molecular pathogenesis to perform better diagnosis and prediction for epithelial ovarian cancer.

Ovarian germ cancer do not occur very often, ~1500 times in UK during 2008 [Jemal et al., 2011], and they are treated different from epithelial ovarian cancers. Two main molecular markers exist to detect ovarian germ cancer: AFP (alpha-fetaprotein), and HCG (human chorionic gonadotrophin). It appears mostly within younger women, and very often are completely cured.

Ovarian cancer has a lack of systems approaches for biomarker discovery and classification based on molecular pathology. These systems approaches have the potential to improve the treatment and management of ovarian cancer in the next decade. In breast cancer (see section 1.3.1 on page 25), a set of biomarkers is well established into a clinical environment. The treatment of ovarian cancer could be improved by finding a similar set of biomarkers.

1.4 Application of systems approach in pathology

During the last decade, more and more *systems* approaches have entered into the molecular biology field [Westerhoff and Palsson, 2004]. This systems approach is widely applied in engineering, software design, and other scientific fields [Hitchins, 2007]. The composition of a system can be a collection of different building blocks. Putting these building blocks together is often called *synthesis*, or a bottom-up approach. Alternatively, a system can be decomposed into smaller building blocks, also called *analysis*, or a top-down approach. Systems biology research can be seen as a pyramid of different building blocks (see figure 1.8). Where the foundations of this pyramid are all the "omics" technologies available for data of genomes, exomes, proteomes, transcriptomes. metabolomes. On top of this "omics" technology, all "ology" disciplines help to formulate more holistic hypotheses of the system under investigation. These two bottom layers help to explain mechanisms that are important in ailments, drugs, and processes in life.

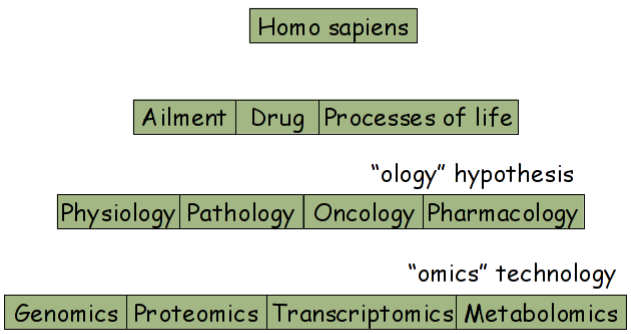


Figure 1.8: A pyramid for system biology illustrates the bottom-up and top-down approaches.

Traditional molecular biology research often happened with a bottom-up approach (reductionism), e.g. pathway analysis. Since more and more "omics" data is publicly available, more top-down approaches can be performed. System biology often occurs in hybrid fashion, e.g. dynamical interactions between pathways are analysed to improve the understanding of cancer biology.

Systems biology becomes more and more the standard approach to perform research in drug discovery (i.e., during 2011: abiraterone, crizotinib, and vemurafenib) [Butcher et al., 2004, Garnett et al., 2012], bio-marker selection (i.e., ER, PR, HER2, and Ki67 for breast cancer) [van't Veer et al., 2005, Faratian and Bartlett, 2008, Faratian et al., 2011], understanding biological mechanisms(i.e., C35 gene expression to indicate tumour invasion) [Faratian et al., 2009].

Reverse engineering [Csete and Doyle, 2002] is a strategy to redesign functionality of an existing system. In software engineering, it is used to obtain the source code from the object code[9] of a program. In systems biology, this term is often used for specific computational techniques applied on "omics" data, e.g. automated reverse engineering ordinary differential equations (ODE) [Bongard and Lipson, 2007], Bayesian networks [Hartemink, 2005], etc.

[9] The object code is the machine code that hardware needs to execute a program.

To conclude this first chapter, a systems approach can enrich conclusions in biology and pathology. This systems approach helps to discover an "omics" footprint of a category of a biological process. This footprint aids to understand the mechanisms of this biological process, and therefore can potentially improve the insights of the heterogeneity; this is required to be able to make more holistic developments that can potentially contribute towards, e.g. a more personalized diagnosis and treatment [Cesario and Marcus, 2011].

The future of systems approaches will require new integrated approaches to examine fairly complex and heterogeneous data sets, e.g. expression data, sequence information, functional annotation, and the literature. Not only data complexity is a major challenge; the growth of sequencing data is beating Moore's law[10] in 2008 [Goldman and Yang, 2008]. As a consequence of this growth, the rate of the cost per genome decreases faster as the rate of the cost per byte in 2008.

[10] Moore's law is used in computer hardware design. It defines that the amount of transistors on integrated circuits doubles every two years.

The bioinformatics and computational biology community proposes a novel approach for data integration: the *gene prioritization strategies* [Moreau and Tranchevent, 2012]. Such a strategy prioritizes genes that are most important for the survival of a patient based upon different "-omics" data resources. Currently, there is an expansion of novel prioritization tools. Each tool uses different data resources, different prior knowledge representations, and different prioritization strategies. The prioritized genes can be checked for their functionality and network interaction [Barabási et al., 2011].

1.5 Layout of the thesis

In the next four chapters, I will first guide you through the methodologies I applied during this research: machine learning techniques and "-omics" technologies. Mainly, I worked on engineering bioinformatics tools that could explain fundamental pathological processes. This novel computational tools are an important facet in the interpretation of heterogeneous data collected with state-of-the-art technologies.

This doctorate is the result of research in two important facets of pathology: (1) biomaker discovery and (2) tumour invasion.

Biomarker discovery The biomarker discovery study is performed for ovarian carcinoma. The data of the Edinburgh Ovarian Cancer Register is used to investigate candidate proteome biomarkers for prognosis. The characterization of the predictability of more traditional clinicopathological measurements and a proteomics profile for prognosis determination is analysed with novel engineered computational tools. The results of these new constructed bioinformatic tools are presented in chapter 3.

These bioinformatics tools were capable to quantify the significance of the set of biomarkers and classify patients if they have a high- or low risk of one-year progression-free and three-year overall survival. Furthermore, I was able to collect a validation data set for an independent group of patients. This allowed me to have hand-on laboratory experience together with the application of high-quality imaging analysis technology. The performance of these computational models after cross-validation and the separate validation data set are at the moment of writing the best found in the literature.

These state-of-the-art bioinformatics tools are also used to construct a biological signature for the histopathological assemblies of the data set. Potentially, they can be applied in many different aspects, e.g. for the validation of various biomarkers in a clinical practice, support engineering of new biomarkers, etc.

The results and the computational methodologies are presented in chapter 3.

Tumour invasion Tumour invasion assays are extremely heterogeneous and pathological characterization of their morphology is important and poorly quantified [Katz et al., 2011]. This tumour invasion study is concentrated on the morphological characterization of tumours. The data collected by collaborators applied state-of-the-art imaging analysis for the collection of morphological measurements that are known to be important for histopathological examination.

The discriminatory capacity of various morphological measurements are investigated. The results are presented in chapter 4.

2
Choice of methodology

> Modernism and postmodernism might be characterised as the two major forces of philosophical thought that have influenced and continue to influence the changes in thinking in research methods. Modernism is associated with the scientific understanding of truth and knowledge, claiming that there is one ultimate, objective truth; and postmodernism relates to the human-centred holistic perspective, maintaining that there are subjective, multiple truths.
>
> *Webster, L. and Mertova, P*

The first part of this chapter introduces the computational methodologies used during this PhD. As presented in the first chapter, I will start with the main generative machine learning approach applied: *Bayesian networks*. Bayesian networks applied to perform *structure learning* will be explained. This will be followed by the discriminative machine learning algorithms, i.e. the traditional *linear models* and *support vector machines (SVMs)*. The final computational approach explained in this chapter is *survival analysis*. This first part will end with the discussion of different machine learning strategies used for the validation of the resulting models.

The second part of this chapter will describe the biological experiments applied in the course of this PhD. *Reverse phase protein arrays (RPPA)* and the *tissue microarrays (TMA)* experiments will be described. I applied TMA technology, in association with collaborators of the Division of Pathology at the University of Edinburgh, for the collection of a validation data set for proteome biomarker validation (see chapter 3).

In this chapter, I will give the background information of the computational modelling performed in chapters 3 and 4. Intially, I

performed Bayesian network analysis to build computational models. In consultation with my supervisors, I have proposed to work with various computational modelling techniques; this techniques are nowadays called machine learning. Machine learning has been applied previously in Bioinformatics [Baldi and Brunak, 2001].

2.1 Computational methodologies

Systems approaches in pathology require computational modelling to quantify their data resources. Since cancer is such a complicated collection of diseases, mathematical modelling aids to understand the underlying patterns. This mathematical modelling is extremely challenging and the predictive capacity inferred form biological data is not always sufficient to construct high quality models [Roberts et al., 2012]. What is modelling and where does a model stands for? A quote of Einstein gives a very good starting point for a usable model:

> "Everything should be made as simple as possible but not simpler".

This applies also to a model, a useful model models a sufficient amount of complexity in as simple as possible way.

In computational modelling one could distinguish two different approaches:

1. *Process-driven approach*:
   - The model is constructed from assumptions, expert knowledge, literature, etc.. The process behind the model is described.
   - Predictions can be made based on *expert knowledge*.
   - Examples of process-driven computational approaches are ordinary differential equations (ODEs), inference over a Bayesian network, etc.
2. *Data-driven approach*:
   - The data is used as the driving-force behind the construction of the model. The model is inferred from experimental data.
   - Predictions are made from *experimental data*.
   - Examples of data-driven computational approaches are linear models, support vector machines, structure learning with Bayesian networks, etc.

In reality process-driven and data-driven approaches are often combined. Combining knowledge from earlier biological research with data from current experiments is a common strategy in computational modelling.

The modelling performed during this PhD will be mainly data-driven: this type of modelling is often called *Data Mining* [Hand et al., 2001]. A potential drawback of these approaches is that they

can be data hungry. Bayesian networks, linear models, survival analysis, etc., are parametrized in such a way that our data resources represent a sufficient sample for modelling.

As described in chapter 1 in page 14, machine learning can be categorized into generative and discriminative machine learning. In the following sections, we will provide an overview of machine learning methodologies applied in different projects. First, a generative machine learning approach will be explained: *Bayesian networks*. This is followed with three discriminative machine learning approaches: (1) *linear models*, (2) *support vector machines (SVMs)*, and (3) *survival analysis*. This chapter will finish with an overview of different validation schemes for machine learning algorithms and performance measures for classification and regression models.

Bayesian networks provide a very strong probabilistic graphical representation of statistical dependencies between a set of random variables. They have one big drawback; their data requirements can be problematic (see section 2.1.1.14 on page 45). Linear models are very often used as a entry point to start data analysis, i.e. they can be used for feature selection in combination with a criterion. Support vector machines (SVMs) are one of the most popular approaches in machine learning. An detailed explaination of SVMs will be provided in this chapter. Censored outputs (i.e., progression-free survival and overall survival) can be modeled with a specific set of computational techniques called *survival analysis*.

One of the most fundamental steps in constructing any computational model is the definition of the preformance measurements. Nowadays, there are various techniques for building a computational model; all techniques can be misused. In order to avoid over- and underfitting problems of a computational model, the performance analysis is more crucial than the applied computational technique. This performance analysis can not always be performed with one measurement and should use preferably a resampling method. Resampling methods can be applied in various ways and attempts to avoid overfitting of mathematical models (e.g., performance measures and model parameters).

All these terminology explained during this chapter will be applied in chapters 3 and 4. I used existing machine learning algorithms, together with a series of performance measurements and resampling methods for constructing novel computational models that have contributed to reveal the system behind pathology.

2.1.1 Bayesian networks

Bayesian statistics have been very popular in the scientific community during the last century. One of the main reason of this success is that prior knowledge can be combined with new data resources. This paradigm resulted in applications in statistical inference [Gelman et al., 2004a], probabilistic graphical models [Koller and Friedman, 2009a], neural networks [Neal, 1995], etc. The Bayesian ap-

proach gained huge interest in systems biology and systems pathology. Systems approaches want to combine prior knowledge with existing and new generated "omics" data set. During this study, a specific Bayesian methodology was applied, the so-called Bayesian networks.

Bayesian networks are probabilistic graphical models, and belong to the category of generative machine learning algorithms (see section 1.1 on page 14). Bayesian networks are conceptually based on Bayes' theorem [Korb and Nicholson, 2004].

$$P(d|s) = \frac{P(s|d).P(d)}{P(s)} \tag{2.1.1.1}$$

- $P(d|s)$: probability of d knowing s.
- $P(s|d)$: likelihood of d resulting in s.
- $P(d)$: probability prior to any evidence d.
- $P(s)$: normalized, so that the conditional probabilities of all hypotheses sum to 1.
- $P(s|d).P(d)$: Joint probability distribution (JPD).

This Bayesian theorem will reoccur in various forms during the explanation of Bayesian networks. Bayesian networks can be applied in various ways [Korb and Nicholson, 2004]. They are applied in different computational modelling approaches (see section 2.1 on page 32):

1. *Data driven approach ("building graphs from data directly")*: these types of algorithms propose graphs directly from a data set. It suggests statistical dependencies between different nodes in a graph. No statistical dependencies are pre-assumed. This example of structure learning algorithms are also called *unsupervised learning*.
2. *Process-driven approach ("updating graphs according known facts")*: these types of inference algorithms proposes graphs starting from *expert knowledge*. The expert knowledge is information provided by an expert (ex.: input graph, certain relationships, certain non-relationships, etc.). It retrieves statistical dependencies according the data set, and creates more probability feedback from the graph used as an input. This example of learning algorithm is also called *semi-supervised learning*.

Probabilistic graphical models (PGM), and Bayesian networks are a combination of basic statistics and computer science [Koller and Friedman, 2009b]. In the following sections, the basic concepts and jargon will be introduced [Spiegelhalter et al., 1993].

Table 2.1: Table of the joint probability distribution (JPD).

r	ras	pfs	JPD
r_1	ras_1	pfs_1	0.025
r_2	ras_1	pfs_1	0.25
r_1	ras_2	pfs_1	0.05
r_2	ras_2	pfs_1	0.275
r_1	ras_3	pfs_1	0.005
r_2	ras_3	pfs_1	0.015
r_1	ras_1	pfs_2	0.001
r_2	ras_1	pfs_2	0.003
r_1	ras_2	pfs_2	0.125
r_2	ras_2	pfs_2	0.002
r_1	ras_3	pfs_2	0.075
r_2	ras_3	pfs_2	0.219

2.1.1.1 Conditioning

Imagine a joint probability distribution over three discrete random variables ($JPD = P(r, ras, pfs)$, see table 2.1 on page 34):

1. Regimen: treatment of a patient with values r_1 and r_2.
2. Ras expression: expression levels of Ras: ras_1, ras_2, and r_3.
3. 1Y-PFS: one year progression-free survival: pfs_1 and pfs_2.

The quantity of parameters is $2 \times 2 \times 3 = 12$, with 11 *independent parameters*. A *JDP* can be conditioned on an observation, e.g. assume we condition on the observation that $ras = ras_2$. This conditional probability distribution is written as $P(r, pfs|ras_2)$, and is computed in two steps: (1) *reduction* (see table 2.2) and (2) *normalization* (see table 2.3). The next operation we can perform is called *marginalization*. If r is marginalized out, we can derive, e.g. the probability for $pfs = pfs_1$.

Table 2.2: Unnormalized probability distribution conditioned on observation $ras = ras_2$.

r	ras	pfs	JPD
r_1	ras_2	pfs_1	0.05
r_2	ras_2	pfs_1	0.275
r_1	ras_2	pfs_2	0.125
r_2	ras_2	pfs_2	0.002

Table 2.3: Normalized probability distribution conditioned on observation $ras = ras_2$.

r	ras	pfs	JPD
r_1	ras_2	pfs_1	0.110
r_2	ras_2	pfs_1	0.608
r_1	ras_2	pfs_2	0.277
r_2	ras_2	pfs_2	0.005

2.1.1.2 Marginalization

Marginalization is a statistical operation on a set of random variables. This operations marginalizes the influence of variable on the resulting probability distribution. In case we marginalize r from table 2.3, our marginalized distribution is the summation of each of the variable states of the subset (see table 2.4).

Table 2.4: Normalized probability distribution conditioned on observation $ras = ras_2$ and marginalized the influence of variable r.

pfs	JPD
pfs_1	0.718
pfs_2	0.282

2.1.1.3 Factor

A factor is a function ($\upsilon(x_1, x_2, \ldots, x_k)$), as every function it has a scope $(x_1, x_2, \ldots, x_k)$. Examples of factors in Bayesian networks are *joint distribution probability (JDP)*, *conditional probability distribution (CPD)*, etc.

In order to introduce Bayesian networks, let's continue with the simple example introduced in the last paragraphs. This example will be used to explain some basic concepts in Bayesian networks (see figure 2.1). These concepts are very important to interpret a Bayesian network.

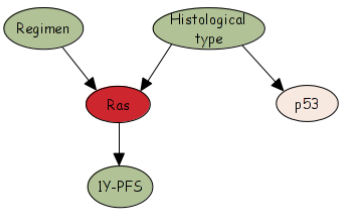


Figure 2.1: A simple example of a Bayesian network to illustrate fundamental concepts.

2.1.1.4 Chain rule for Bayesian networks

All the conditional distribution probabilities (*CPD*) formulate the joint distribution probability (*JPD*) by the application of the *chain rule for Bayesian networks* (see figure 2.2):

$$P(R, H, Ras, p53, PFS) = P(R)P(H)P(Ras|R, H)P(p53|H)P(PFS|Ras) \tag{2.1.1.2}$$

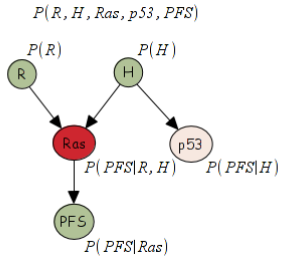


Figure 2.2: A simple example of a Bayesian network to illustrate the conditional distribution probabilities (*CPD*) that are part of the chain rule.

This allows us to compute the joint probability of any combination of values of our system under investigation, e.g.

$P(r_1, h_2, ras_2, p53_1, pfs_2) = 0.27 \times 0.21 \times 0.31 \times 0.23 \times 0.47 = 0.0019$

This defines a general definition for Bayesian networks. A Bayesian network is a directed acyclic graph (DAG) whose nodes represent

Table 2.5: Conditional probability table for Regimen.

r_1	r_2
0.27	0.73

the random variables $\chi = (X_1, X_2, \ldots, X_n)$ of our joint probability distribution; the edges, are the statistical dependencies among those random variables, represent the conditional probability distribution for each node ($CPD(X_i) = P(X_i|Par_G(X_i))$). A DAG means that there are no loops in the network. If we apply the chain rule for Bayesian networks, then we can write the joint probability distribution (JPD). This JPD is a factor product of the condition probability distribution, of our Bayesian network as:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|Par_G(X_i)) \tag{2.1.1.3}$$

This chain rule for a Bayesian network is a very important formulation. An alternative interpretation of a Bayesian network is given by a set of conditional independencies. The concept of conditional independency and how it is encapsulated in a Bayesian network will be explained in the following sections.

Table 2.6: Conditional probability table for histological type.

h_1	h_2
0.79	0.21

2.1.1.5 Independence

If two random variables, X and Y, are independent, we can write the following: $X, Y, P \vDash X \perp Y$

$$P(X, Y) = P(X)P(Y) \tag{2.1.1.4}$$
$$P(X|Y) = P(X) \tag{2.1.1.5}$$
$$P(Y|X) = P(Y) \tag{2.1.1.6}$$

2.1.1.6 Conditional independence

A set of three random variables X, Y, and Z where X and Y are conditional independent given Z: $P \vDash (X \perp Y|Z)$

$$P(X, Y|Z) = P(X|Z)P(Y|Z) \tag{2.1.1.7}$$
$$P(X|Y, Z) = P(X|Z) \tag{2.1.1.8}$$
$$P(Y|X, Z) = P(Y|Z) \tag{2.1.1.9}$$
$$P(X, Y, Z) \propto \upsilon_1(X, Y)\upsilon_2(Y, Z) \tag{2.1.1.10}$$

Table 2.7: Conditional probability table for Ras.

		ras_1	ras_2	ras_3
r_1	h_2	0.1	0.68	0.22
r_1	h_2	0.51	0.31	0.18
r_2	h_2	0.28	0.23	0.49
r_2	h_2	0.13	0.15	0.72

The conditional independencies in a Bayesian network are often explained by an *active trail*. In the next section, an active trail will be explained together with our simple example from the previous sections.

Table 2.8: Conditional probability table for p53.

	$p53_1$	$p53_2$	$p53_3$
h_1	0.22	0.48	0.3
h_2	0.23	0.02	0.71

2.1.1.7 Active trail

A very important, and at the same time an often misunderstood, concept is to understand how Bayesian networks represent statistical dependencies in the graph. In the following paragraphs an example illustrates how Bayesian networks allow you to reason over the directed graph. A very fundamental question to ask is: "Is node X_i independent on X_j?". This type of reasoning over a Bayesian network asks for the introduction of some novel concepts.

Table 2.9: Conditional probability table for 1Y-PFS.

	pfs_1	pfs_2
ras_1	0.71	0.29
ras_2	0.53	0.47
ras_3	0.12	0.88

Reasoning on a Bayesian network can be performed in two manners: (1) without a set of evidence nodes and (2) with a set of evidence nodes. Reasoning is often performed with an *active trail*. An active trail in a Bayesian network means that node X_i has a statistical dependency on node X_j.

A trail is active in a Bayesian network if:

1. *without a set of evidence nodes*[1]:
   An trail ($X_1 \longleftrightarrow X_2 \longleftrightarrow \ldots \longleftrightarrow X_k$) in a Bayesian network is active if the connected nodes in a Bayesian network contain no *v-structure* ($X_{i-1} \longrightarrow X_i \longleftarrow X_{i+1}$, see figure 2.3 on page 37).
2. *with a set of evidence nodes* (Ξ):
   A trail ($X_1 \longleftrightarrow X_2 \longleftrightarrow \ldots \longleftrightarrow X_k$) in a Bayesian network is active given the set of evidence nodes (Ξ) if:
   - for any v-structure ($X_{i-1} \longrightarrow X_i \longleftarrow X_{i+1}$) X_i, or any of its descendants $\in \Xi$
   - any other $X_{1 \to k} \in \Xi$

A v-structure in a Bayesian network (see figure 2.3) without a set of evidences nodes blocks an active trail. A trail with a v-structure can only be activated if the node *Ras* or its descendant *PFS* is an element of the set of evidence nodes (Ξ).

A parent node (X_j) and a child node X_i in a Bayesian network are statistical dependent. Obviously, these statistical dependencies are bidirectional; this can be misleading since the arrow points only in the direction of the child ($X_j \longrightarrow X_i$).

In a longer trail between two nodes in the example Bayesian network, node *R* is connected with node *PFS* (see figure 2.4). In case there is no evidence set, this trail is active. Alternatively, in case node *Ras* is in the evidence set (Ξ), then the trail is no longer active. Such top-down trail is also called *causal reasoning*.

Analogously, when the trail between two nodes in the example Bayesian network node *PFS* and *R* are connected (see figure 2.5). If there is no evidence set (Ξ) then the trail is active. It is only in the case node *Ras* is an element of the evidence set (Ξ) that the trail is not active. This type of bottom-up reasoning is often called *evidential reasoning*.

A trail in a Bayesian network where a parent node has two child nodes ($X_{i-1} \longleftarrow X_i \longrightarrow X_{i+1}$, see figure 2.6) is active if there is no set of evidence nodes. In case the parent node (X_i) is an element of the set of evidence nodes (Ξ), the trail is not active (inter-causal reasoning).

In case a trail, $p53 \longleftarrow H \longrightarrow Ras \longleftarrow R$, is activated, *H* may not be part of the evidence set (Ξ) and *Ras*, or its descendant (*PFS*), must be part of its evidence set (Ξ; see figure 2.7).

2.1.1.8 *d-separation*

Active trails in a Bayesian network are important to retrieve the conditional independencies [Geiger et al., 1990]. Three nodes in a

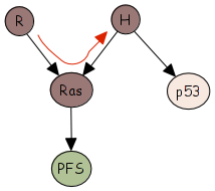


Figure 2.3: The Bayesian network example were the *v-structure* is indicated.

[1] An evidence node is a node that we have evidence of its probability distribution. If data is collected it can calculation.

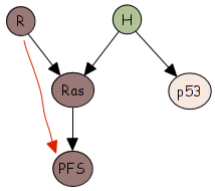


Figure 2.4: Statistical dependencies in a Bayesian network a top-down information flow (causal reasoning).

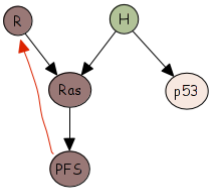


Figure 2.5: Statistical dependencies in a Bayesian network a bottom-up information flow (evidential reasoning).

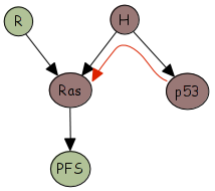


Figure 2.6: Statistical dependencies in a Bayesian network with a $X_{i-1} \longleftarrow X_i \longrightarrow X_{i+1}$ (inter-causal reasoning).

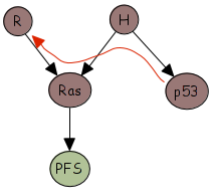


Figure 2.7: A trail in a Bayesian network that combines different types of reasoning.

Bayesian network, X_i, X_j, and X_k, where X_k d-separate X_i from X_j, if and only if there is no active trail between X_i and X_j given X_k.

It is important to understand that factorization and conditional independency are both represented into a Bayesian network. From a Bayesian network we can write the joint probability that represents all the directed connections of the acyclic graph and all conditional independencies. This set of conditional independencies, or d-separations (direction-dependent separation), are often called an *I-map*.

So far we introduced Bayesian networks without any time related information. Therefore this category of Bayesian networks are called *static Bayesian networks*.

2.1.1.9 Static Bayesian network

Static Bayesian networks are restricted to be *directed acyclic graphs (DAG)*. A static Bayesian network encodes a Joint Probability Distribution (JPD) over a set of discrete variables ($\chi = \{X_1, X_2, \ldots, X_n\}$) [Heckerman et al., 1995].

A Bayesian network is mathematically represented by a graph (G) and a set of parameters (Θ) which describe the probability of variables taking on each of their discrete values [Heckerman et al., 1995].

The Bayesian network of χ can be represented as:

$$BN_{static \rightarrow \chi} =< G, \Theta > \tag{2.1.1.11}$$

Graph (G) The *nodes* in the graph are the random variables ($\chi = \{X_1, X_2, \ldots, X_n\}$). The *directed links* represent the statistical conditional dependencies (X_j on X_i) [Yu, 2005].

Set of variables (Θ) For each random variable of the probability distribution we can write:

$$\Theta_{x_i|Par(X_i)} = P(X_i = x_i|Par(x_i) = par(X_i) \tag{2.1.1.12}$$

The probability of X_i taking on the value x_i given its parents $Pa(X_i)$ having the values in a particular instantiation of the parents, $par(X_i)$, for all x_i and $par(X_i)$ [Yu, 2005].

Static Bayesian networks have specific limitations. As mentioned earlier, the graph needs to be acyclic. A unique JPD for a BN can have several different equivalent factorings (only directions of some links differ). Dynamic Bayesian networks can bypass some of these limitations [Yu, 2005], e.g. in case we have time series data of the system under investigation, we can obtain feedback loops.

2.1.1.10 Temporal models

Temporal models are computational representations of *temporal trajectory distributions*. Often the time is discretized, such data sets are called *time series*. Time series have often a granularity (Δ), where

each variable at a time frame t is written as $X_{(t)}$, and time series between t_i and t_j: $X_{(i:j)} = \{X_i, X_{i+1}, \ldots, X_j\}$.

A temporal model will describe a distribution over trajectories, this formulates as $P(X_{i:j})$. An important assumption for temporal models, in order to compactify this probabilistic distribution, is called the *Markov assumption*.

Markov assumption Applying the chain rule of probabilities allows us to write

$$P(X_{(0:T)}) = P(X_{(0)}) \prod_{t=0}^{T-1} P(X_{(t+1)}|X_{(0:t)}) \tag{2.1.1.13}$$

$$X_{(t+1)} \perp X_{(0:t)}|X_{(t)} \tag{2.1.1.14}$$

$$P\left(X_{(0:T)}\right) = P(X_{(0)}) \prod_{t=0}^{T-1} P(X_{(t+1)}|X_{(t)}) \tag{2.1.1.15}$$

The Markov assumption is a *"forgetting assumption"*: we derive the next state ($X_{(t+1)}$) based on the current state ($X_{(t)}$), and forget about the past ($X_{(0:t)}$). This Markov assumption might be violated in certain applications, therefore there are two main strategies to make this Markov assumption true. First strategy is to add more information about the state during each step in trajectory. The second strategy is to include more steps back into history where the next step will be based on. In this case we define it as a n-order Markov assumption, where n indicates the number of steps taken into account.

Time invariance assumption A template probability model that follows the Markov assumption, can also follow the time invariance assumption. This assumption restricts the dynamics of the model during the trajectory. These dynamics are assumed to be equal between two successive time points. Such a model is also called *stationary* or *homogeneous* [Koller and Friedman, 2009b].

$$P\left(X_{(t+1)}|X_{(t)}\right) = P\left(X'|X\right) \tag{2.1.1.16}$$

2.1.1.11 Dynamic Bayesian network (DBN)

Dynamic Bayesian networks include the dimension of time. Often a *first order Markov assumption* is used. This implies that variables at one time slice are considered to be affected only by those in the immediately previous time slice [Yu, 2005].

Such a DBN is a graphical representation of a joint probability distribution over χ': set of discrete random variables X_i measured at times t and $t + \Delta t$ [Heckerman et al., 1995]:

$$\chi' = \{X_{1_{(t)}}, X_{2_{(t)}}, \ldots, X_{n_{(t)}}, X_{1_{(t+\Delta t)}}, X_{2_{(t+\Delta t)}}, \ldots, X_{n_{(t+\Delta t)}}\} \tag{2.1.1.17}$$

The Bayesian network of χ' can be represented as [Yu, 2005]:

$$BN_{dynamic \rightarrow \chi'} =< G, \Theta > \tag{2.1.1.18}$$

Graph (G) Links are only possible forward in time ($X_i(t) \longrightarrow X_j(t + \Delta t)$) and all variables have links to themselves ($X_i(t) \longrightarrow X_i(t + \Delta t)$) [Yu, 2005].

$$P(X_{(t+\Delta t)}|X_{(0)}, X_{(1)}, X_{(2)}, \ldots, X_{(t)}) = P(X_{(t+\Delta t)}|X_{(t)}) \tag{2.1.1.19}$$

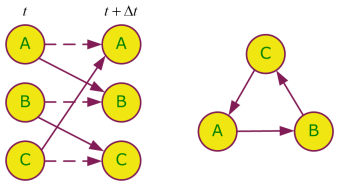


Figure 2.8: DBN representation for an underlying causal network with loop.

Set of variables (Θ) As above a collection of variables consists for all $X_i(t + \Delta t)$ in χ' [Heckerman et al., 1995]:

$$\Theta_{x_i(t+\Delta t)|Par(X_{i_{(t+\Delta t)}})} \tag{2.1.1.20}$$

This formulation of a dynamic Bayesian network is also called *2-time-slice Bayesian network (2TBN)*. In the previous sections we provided the theoretical background to perform inference and reasoning in a Bayesian network, and explained the difference between static and dynamic Bayesian networks. During this thesis Bayesian networks are often applied for *structure learning*. Structure learning algorithms for Bayesian networks will be explained in the following sections.

2.1.1.12 Structure learning

Structure learning of Bayesian networks can be performed in two manners: (1) *construct an I-map based on conditional independence tests* [Cheng et al., 1997, Daly et al., 2009] and (2) *search and score structure* [Heckerman et al., 1995, Heckerman, 1996, de Campos, 2006, Cowell, 2001][2]. In the course of this study I concentrated on the search and score structure approach. In the following paragraphs, I will provide an intuition for the derivation of a scoring metric for the structure learning of a Bayesian network[3].

[2] **R package**:
The following R packages support various Bayesian network analysis:

- bnlearn [Scutari, 2010]: various different constraint independence (CI) tests, and scoring metrices for Bayesian network analysis are available.
- deal [Bottcher and Dethlefsen., 2009]: Bayesian network learning with discrete and continuous variables.

[3] **Java framework**:
The following Java framework supports Bayesian network analysis:

- BANJO [Hartemink, 2005]: Bayesian Network Inference with Java Objects.

Scoring metric How big is the probability that a graph G can explain the data in D? There are two general approaches to score a graph: (1) *maximum likelihood scores* and (2) *Bayesian scores* [Cooper and Herskovits, 1992]. In this PhD, the main focus was on Bayesian scores for structure learning of Bayesian networks. A Bayesian score can be directly derived from the Bayes' rule (see equation 2.1.1.1 on page 34) [Heckerman, 1996]:

$$P(G|D) = \frac{P(D|G)P(G)}{P(D)} \tag{2.1.1.21}$$

$$P(G|D) \propto P(D|G)P(G) \tag{2.1.1.22}$$

$$Score(G : D) = log(P(G|D)) = log(P(D|G) + log(P(G))) \tag{2.1.1.23}$$

In the Bayes' rule the denominator is seen as a *normalization term*, and will therefore not be informative about the structure. The *prior*

distribution over all the graphs, $P(G)$, specifies the preference over certain graphs and is often chosen *uniform*. Designing a Bayesian scoring metric is equivalent to the *marginal likelihood*[4] or in some literature called *evidence*, $P(D|G)$, which can be written as:

$$P(D|G) = \int_{\theta_G} P(D|\theta_G, G)P(\theta_G|G)d\theta_G \tag{2.1.1.24}$$

Computation of a Bayesian scoring metric is done by marginalizing out (see section 2.1.1.2 on page 35) the parameters of the graph θ_G, $P(\theta_G|G)$ is the prior distribution of the parameters of the graph, and $P(D|\theta_G)$ is likelihood of the data given the Bayesian network structure parameters. The marginal likelihood follows the *Occam's Razor principle* [Domingos, 1999], meaning that it favors less complicated structures; this property must be understood as a consequence of more complex structures contain more graph parameters (θ_G), which can not contain more probability mass as instinctively available in the data since the probability constraint sums to one.

If we assume that all graph parameters are independent (see equation 2.1.1.5 on page 36), we can write:

$$P(D|G) = \prod_{i=1}^{n} \int_{\theta_i} P(X_i|Par_G(X_i), \theta_i)P(\theta_i) \tag{2.1.1.25}$$

There exist many different Bayesian scoring metrices. Some of the scoring metrices are based upon *discrete variables*, e.g. Bayesian Dirichlet equivalent, K2, etc., others are based upon *continuous variables*, e.g. Bayesian Gaussian equivalent [Heckerman and Geiger, 1995, Nodelman et al., 2002], Bayesian information criterion score, etc., and there exist also hybrid [Bottcher and Dethlefsen., 2009] Bayesian scoring metrices. During this PhD, mainly Bayesian scoring metrices for discrete variables are applied. In the following sections, the Bayesian Dirichlet equivalent (BDe) score will be explained; the BDe score is one of the most used scoring metrices for structure learning in Bayesian networks. An alternative Bayesian scoring metric is the *Bayesian Information Criterion (BIC)*.

Both scoring metrics involve the generation of a *Conditional Probability Table (CPT)* for each node.

Conditional probability distribution (CPD) For discrete nodes the local conditional probability distribution (CPD) is a multinomial distribution, which results in conditional probability tables (CPT) for each node. The conditional probability table (CPT) stores the probabilities estimated from all combinations of parent-child values extracted from the discretized data $\theta_{ijk} = P(x_i = k|Par(x_i) = j)$ [Heckerman, 1996].

A possible solution to solve equation 2.1.1.25 for discrete nodes is based on three assumptions: (1) *global parameter independence*, (2) *local parameter independence*, and (3) *likelihood equivalence*. Global parameter independence defines that a node in a Bayesian network (X_i) is independent of all the other nodes given its parents ($Par(X_i)$).

[4] Marginal likelihood is different from the maximum likelihood score (see section 2.1.1.12). The maximum likelihood examine the maximum of the likelihood function of the data given the network structure, whereas the marginal likelihood calculates the average, based on $P(\theta_G|G)$, of the same likelihood function. This is a main cause of the risk for over-fitting when maximum likelihood score is applied.

We can write for the parameters θ_i of node X_i:

$$P(\theta) = \prod_{i=1}^{n} P(\theta_i) \tag{2.1.1.26}$$

$$\theta_i = \left\{\theta_{ijk}; j = 1, \ldots, q_i, k = 1, \ldots, r_i\right\} \tag{2.1.1.27}$$

Global parameter independence specifies that the Bayesian score is a *decomposable score*; this score can be decomposed into different terms, and is a very important property for a heuristic search algorithm for finding the highest score.

Local parameter independence defines that the parameters of each node (X_i) given a parent are independent of the parameters for the same node given other parent(s).

$$P(\theta) = \prod_{i=1}^{q_i} \theta_{ij} \tag{2.1.1.28}$$

$$\theta_{ij} = \left\{\theta_{ijk}; k = 1, \ldots, r_i\right\} \tag{2.1.1.29}$$

Likelihood equivalence defines that two network structures with equal I-maps result into equal marginal likelihood ($P(D|G_1) = P(D|G_2)$). In case these three assumptions are not violated[5], the prior distribution can be a *Dirichlet prior*.

[5] These three assumptions can be violated is the global- and local independencies are not true for the score we want to formulate, or equal I-maps should not represent an equal score.

Dirichlet prior Each CPD of a variable in the Bayesian network is a multinomial ($P(X_i|Par(X_i) = j) = \theta_{ij}$) with r_i possible discrete values. The Dirichlet prior is defined as:

$$\theta_{ij} = Dir(\alpha_{ij1}, \alpha_{ij2}, \ldots, \alpha_{ijk}) \tag{2.1.1.30}$$

$$P(\theta_{ij}|\alpha_{ij}) = \frac{\prod_{k=1}^{r_i} \Gamma(\alpha_{ijk})}{\Gamma(\sum_{k}^{r_i} \alpha_{ijk})} \prod_{k=1}^{r_i} \theta_{ijk}^{\alpha_{ijk}-1} \tag{2.1.1.31}$$

The Dirichlet prior distribution is a conjugate[6] for the Bayesian score. Now we can derive the Bayesian Dirichlet equivalent (BDe) scoring metric.

[6] If the posterior distribution is of the same family as the prior distribution, the prior and posterior are called *conjugate* distributions [Gelman et al., 2004b].

Bayesian Dirichlet equivalent (BDe) scoring metric The *BDe score* captures the full Bayesian posterior probability $P(G|D)$. In this metric, the prior over graphs needs to be specified (usually the uniform prior is applied) and *the prior over parameters is Dirichlet*, a distribution over multinomial distributions describing the *conditional dependency of each variable in the network*.

$$BDe(G:D) = P(D|G) = \prod_{i=1}^{n} \left(\prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + n_{ij})} \left(\prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + n_{ijk})}{\Gamma(\alpha_{ijk})} \right) \right) \tag{2.1.1.32}$$

$$\log(BDe(G:D)) = \sum_{i=1}^{n}\sum_{j=1}^{q_i}\left(\log\Gamma(\alpha_{ij}) - \log\Gamma(\alpha_{ij}+N_{ij}) + \sum_{k=1}^{r_i}\left(\log\Gamma(\alpha_{ijk}+n_{ijk}) - \log\Gamma(\alpha_{ijk})\right)\right) \tag{2.1.1.33}$$

The *Gamma function* is calculated for $n \in \mathbb{N}_0^+$:

$$\Gamma(n) = (n-1)! \tag{2.1.1.34}$$

The following table 2.10 provides an explanation of the most important variables required for the computation of the Bayesian Dirichlet equivalent.

Variable	description
$\prod_{i=1}^{n}$	n: number of variables in the data set.
$\prod_{j=1}^{q_i}$	q_i: number of joint parent states of a child X_i ($q_i = 1$ if $Par_i = 0$).
α	Equivalent sample size (*ess*) [Silander et al., 2007], expresses prior knowledge ($\alpha = 0$ no prior knowledge).
α_{ij}	Specific value for each BD variant, i.e. BDeu: $\alpha_{ij} = \frac{\alpha}{q_i}$
n_{ij}	Quantity of times for a variable X_i that the parents are in joint state j, regardless of the state of X_i. Sum of cases of a particular value ($n_{ij} = \sum_{k=1}^{r_i} n_{ijk}$).
$\prod_{k=1}^{r_i}$	quantity of different states a variable X_i takes over the complete data set (quantity of states of a child).
α_{ijk}	Pseudo counts (hyperparameters), i.e. BDeu variant: $\alpha_{ijk} = \frac{\alpha}{r_i \cdot q_i}$)
n_{ijk}	Quantity of times variable X_i is in state k, while its parents are in joint state j ($n_{ijk} = P(x_i = k \mid Par(x_i) = j)$). Knowing the state of a child ($state_{child}$), how many times the parents have a particular state?

Table 2.10: Different variables and corresponding description used for the computation of Bayesian Dirichlet equivalent (BDe) scoring metric.

BD score variants Most of the BD variants define a uninformative prior; in Bayesian statistics it is better to use an uniform or uninformative prior distribution instead of an incorrect prior distribution[7] [Gelman et al., 2004b].

[7] Bayesian statistics can produce unmeaningful results in case the prior distribution is not defined properly. An uniform of uninformative prior distribution is often preferred for structure learning with Bayesian networks.

There are a few variants of BD scores. Some of these variants violate the equivalent likelihood assumption, i.e. K2 score, others do keep this assumption, i.e. DBe and DBeu score.

The K2 score has an uninformative prior with an equivalent sample size (α) defined as:

$$\alpha_{ij} = r_i \tag{2.1.1.35}$$

$$\alpha_{ijk} = 1 \tag{2.1.1.36}$$

$$K2(G:D) = P(D|G) = \prod_{i=1}^{n}\left(\prod_{j=1}^{q_i}\frac{(r_i-1)!}{(n_{ij}+r_i-1)!}\left(\prod_{k=1}^{r_i}(n_{ijk})!\right)\right) \tag{2.1.1.37}$$

The BDeu (u stands for uniform distribution) assigns a uniform distribution to the joint distribution. The pseudo counts are set in such a manner that each network is equally likely.

$$\alpha_{ij} = \frac{\alpha}{q_i} \tag{2.1.1.38}$$

$$\alpha_{ijk} = \frac{\alpha}{r_i q_i} \tag{2.1.1.39}$$

2.1.1.13 *Data discretization*

Data discretization is an algorithm that transforms continuous variables into discrete values [Butterworth et al., 2004]. Discretization of continuous variables allow us to apply structure learning with scoring metrices for discrete variables. Data discretization influences substantially the resulting Bayesian network [Steck and Jaakkola, 2006].

Assume n continuous variables in the domain of interest:

$$Y = (Y_1, Y_2, \ldots, Y_n) \tag{2.1.1.40}$$

In the following paragraphs, a number of these policies will be illustrated.

Interval or range discretization A n-way interval discretization for a data range between $x \ldots y (y > x)$:

$$range_1 : 0 \ldots \frac{y-x}{n} \tag{2.1.1.41}$$

$$range_2 : \frac{y-x}{n} \ldots 2.\frac{y-x}{n} \tag{2.1.1.42}$$

$$\ldots$$

$$range_{n-1} : (n-2).\frac{y-x}{n} \ldots (n-1).\frac{y-x}{n} \tag{2.1.1.43}$$

$$range_n : (n-1).\frac{y-x}{n} \ldots n.\frac{y-x}{n} \tag{2.1.1.44}$$

Quantile discretization The major characteristic of *quantile discretization* is to have equal number of each n-way quantile discretization levels [Steck and Jaakkola, 2006].

Fayyad-Irani's discretization This discretization policy is often used in Data Mining and Machine Learning [Fayyad and Irani, 1993]. It applies an *entropy minimization heuristic* in the discretization algorithm [Fayyad and Irani, 1993][8].

Hartemink's pairwise mutual information discretization This discretization policy attempts to minimize the total pairwise information loss [Hartemink et al., 2001, Grzegorczyk, 2006][9]. Mutual information is a measurement to define the *dependency* or *distance* between

[8] **Python package**:
The following Python package to support different discretization policies:

- orange [Janez Demsar and Curk, 2004]: interval, quantile, and Fayyad-Irani's discretization policies are available.

[9] **R package**:
The following R package to support different discretization policies:

- bnlearn [Scutari, 2010]: interval, quantile, and Hartemink's pairwise mutual information discretization policies are available.

variables. Many frameworks quantify the linear dependency between variables, mutual information provides a *general* measure of dependencies among variables [Steuer et al., 2002]. We can also say that mutual information quantifies the stochastic dependence, or the degree of predictability, between two variables [Hausser, 2006]. The mutual information between X and Y is defined by [MacKay, 2004]:

$$MI(X,Y) = \sum_{i=1}^{M}\sum_{j=1}^{M} P(X=i,Y=j)\log_2\frac{P(X=i,Y=j)}{P(X=i)P(Y=j)} \quad (2.1.1.45)$$

From equation 2.1.1.45, we can derive that the mutual information is the ratio between the joint distribution $P(X=i,Y=j)$ and the product marginal distributions [Hausser, 2006]. In case the joint distribution and the product of marginals are equal, both variables are stochastically independent.

2.1.1.14 *Data requirements*

Bayesian networks tend to be data intensive[10] [Yu, 2005]. In order to avoid *false positives* in the graph, sufficient data samples must be available. The following table gives an overview of the quantity of data needed for structure learning with the BDe scoring metric:

[10] Data intensive means the a rather high quantity of data samples are required compared to other other compared to other computationsl methods.

Number of discretization levels	Number of samples: 1 parent - child relationship	Number of samples: 2 parent - child relationships	Number of samples: 3 parent - child relationships	Number of samples needed: 4 parent - child relationships	Number of samples needed: 5 parent - child relationships	Number of samples needed: 6 parent - child relationships
2	6	12	24	48	96	192
3	15	45	135	405	1215	3645
4	28	112	448	1792	7168	28672

Table 2.11: Overview of the number of samples needed in order to avoid false positives in function of the quantity of parent - child relationships and the number of discretization levels ($\alpha = 2$).

Number of discretization levels	Number of samples: 1 parent - child relationship	Number of samples: 2 parent - child relationships	Number of samples: 3 parent - child relationships	Number of samples needed: 4 parent - child relationships	Number of samples needed: 5 parent - child relationships	Number of samples needed: 6 parent - child relationships
2	2.27	6.05	15.13	36.31	84.72	193.66
3	8.01	34.18	132.99	409.3	1744.86	6056.47
4	17.35	101.43	533.83	2647.77	12640.98	8763.46

Table 2.12: Overview of the number of samples needed in order to retrieve the minimal number of samples needed to find parents (*alpha* = 2).

2.1.1.15 *Heuristic search methods*

Identifying the highest scoring graph is a *Non-deterministically Polynomial (NP)* complete problem [Chickering, 1996, Korb and Nicholson, 2004]. This is a category of complexity theory problems with an inherent intractability. Heuristic search algorithms are used to improve results found in a more reasonable time.

Simulated Annealing (SA) [Cerny, 1985] and *Greedy search with random restarts* [Chickering, 2003] are both applied in order to verify

that the modelling results are independent on the applied search algorithm [Yu et al., 2004].

Recently, there are search algorithms applied for various problems: Markov Chain Monte Carlo (MCMC) [Grzegorczyk and Husmeier, 2008], ant colony optimization (ACO) [Daly and Shen, 2009], max-min hill-climbing [Tsamardinos et al., 2006] etc. Many of these algorithms have been implemented for structure learning of Bayesian networks.

2.1.1.16 *Model averaging*

Model averaging is applied for structure learning with Bayesian networks since one best network structure learned might have missed important statistical dependencies among the random variables. Bayesian model averaging (BMA) [Madigan and Raftery, 1994, Hoeting et al., 1999] provides a framework to avoid over-fitting of the selected model.

Since the modelling technique reflects the correctness of a certain network for describing a data set by one score, we can loose some important patterns. We use model averaging to capture more patterns from the data set as an attempt to capture edges from other high-scoring networks [Hartemink et al., 2002].

Model averaging of a Bayesian network can be performed over, e.g. best 100 networks, formulated as [Hartemink et al., 2002]:

$$p(E_{XY}|D) \approx \frac{\sum_{i=1}^{N} 1_{XY}(S_i).e^{BSM(S_i)}}{\sum_{i=1}^{N} e^{BSM(S_i)}} \tag{2.1.1.46}$$

The symbols in these formula represent:

- N: number of best graphs according the BDe score.
- E_{XY}: edge between variable X and Y.
- $1_{XY}(S_i)$: is equal to 1 if and only if edge E_{XY} is part of network S_i
- $BSM(S_i)$: the Bayesian scoring metric for graph S_i.

2.1.1.17 *Influence score*

The influence score (Θ_{ijk}) for Bayesian networks informs the *sign* and *magnitude of influence* of a specific edge in the graph [Yu et al., 2004]. The influence score of a statistical dependency between a parent- and a child node in the Bayesian network is derived from conditional probability using a counting and voting mechanism described in [Hartemink et al., 2001]. Its meaning is described in the following paragraphs.

Sign The sign of the influence score expresses the following information [Hartemink et al., 2001]:

- *positive (+) from X to Y*: higher values of parent node X will bias the distribution of child node Y higher.
- *negative (-) from X to Y*: higher values of parent node X will bias the distribution of child node Y lower.
- *zero*: there is no monotonic influence from X to Y, i.e. U- or hump shaped relationships.

Magnitude The magnitude specifies the strength of the influence of a parent node on a child node [Yu et al., 2004].

In the following section, linear models will be explained.

2.1.2 Linear models

One of the most widely used discriminative and supervised machine learning algorithms are *linear models*. It is often used as a starting point for more complicated and nonlinear models. In the next sections the general- and generalized linear models are introduced.

Throughout the following sections, the data will have following characteristics. The data set with m samples is constructed of two parts: (1) *n explanatory variables*, also called *features* (x) and (2) the *response- or output variable* (y)[11]:

[11] $x_i^{(j)}$: is the value of the ith feature and the jth data sample.

$$\forall j \in \{1, 2, \ldots, m\} : d^{(j)} = \left\{x^{(j)}, y^{(j)}\right\} \tag{2.1.2.1}$$

$$\forall i \in \{1, 2, \ldots, n\} : x \in \mathbb{R}^{n \times m} \tag{2.1.2.2}$$

$$y \in \mathbb{R}^1 \tag{2.1.2.3}$$

A linear regression analysis looks for a linear relationship between explanatory variables, also called features, and a response- or output variable:

1. *Explanatory variables (features; x)* can be *continuous* or *ordinal categorical*. Categorical is also called *nominal*. It means that a variable has two of more categorical values. Ordinal specified that their exists a clear ordering of the variables.
2. *Response variable (output variable; y)* is the output variable that we want to predict. In a regression model the response variable is *continuous*, and in a classification model the response variable is categorical.

In the case of multiple explanatory variables, a linear model is often called *Multiple Linear Regression (MLR)*.

A regression analysis has two levels of interpretation: (1) *basic level* interprets the association retrieval between the explanatory and response variables and (2) *sophisticated level* specifies the prediction of response variables based on the explanatory variables.

A linear model constructs an hypothesis that a set of parameters (θ) explain the relationship among the features and the output variable. This relationship can be written as:

$$h_\theta = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n \tag{2.1.2.4}$$

A cost function is a performance measure that quantifies the error of a linear model constructed with a set of parameters ($\theta = \{\theta_0, \theta_1, \theta_2, \ldots, \theta_n\}$). In the case of the data set has m data points and n features we can write the cost function as[12]:

[12] The cost function used is the mean squared error (MSE)

$$J(\theta_0, \theta_1, \theta_2, \ldots, \theta_n) = \frac{1}{2m} \sum_{j=1}^{m} \left(h_\theta(x^{(j)}) - y^{(j)} \right)^2 \tag{2.1.2.5}$$

The goal of constructing a linear model is to minimize the cost function for a set of parameters (θ) given the data set. A linear model can be constructed with different statistical software packages: R, Octave, Python, etc. Now let's discuss general linear models and the interpretation of the output provided by many statistical packages.

2.1.2.1 *General Linear Model (GLM)*

The general linear model (GLM) framework is the tool to construct ordinary linear models [Grafen and Hails, 2006][13].

[13] **R package**:
The following R functions to support general linear models:
- lm: linear model function that fits parameters θ.

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n + \epsilon \tag{2.1.2.6}$$

Typically, the set of parameters ($\theta = (\theta_0, \theta_1, \theta_2, \ldots, \theta_n)$) are computed to minimize the error distribution of the GLM. This error distribution, $\epsilon = N_n(0, \sigma^2 I)$, is a multivariate normal distribution. Linear models with different error distributions are called *generalized linear models (GeLM)* [Olsson, 2002].

GLM's can handle two types of features: categorical- and continuous features [Grafen and Hails, 2006]. An analysis with categorical features results in retrieval of difference between *mean values* between the different categories and an analysis with continuous features retrieves the linear relationships with the response variable. An analysis with only categorical variables is called the *t-test* and *analysis of covariance*, an analysis with only continuous variables is called *regression* or *multiple regression*, and an analysis with a combination of categorical and continuous is called *analysis of covariance* [Grafen and Hails, 2006].

[14] **R package**:
The following R functions to support contrast schemes for categorical variables:
- factor: a factor type definition indicates a categorical variable in R.
- C: allows specific contrast representations. e.g. treatment, helmert, sum, and poly.
- contr.: allows specific contrast representations. e.g. contr.treatment, contr.helmert, contr.sum, and contr.poly.

Categorical variables Categorical variables need special attention if they are added into a linear model [Serlin and Levin, 1985]. Intuitively, a categorical variable of k levels can be represented by $k-1$ columns in a matrix. This representation is performed by different coding schemes, also called *contrasts*. There are many different contrast procedures available, e.g. *treatment* or also called *dummy*, *Helmert*, *sum*, and *poly* [Harrell, 2001, Harrell, 2012b][14]. A detailed

description of each of these coding schemes is documented in a report written by Sundström [Sundström, 2010].

The coding schemes of this categorical data help with the interpretation of the effect of an individual coding variable, but it does not change the overall effect of a set of coding variables on the linear model fit.

Output The following paragraphs explain the main output diagnostics that a statistical software package provides for a linear model, i.e. ANOVA table, coefficient table, confidence- and prediction intervals, etc. These output diagnostics are very similar for generalized linear models and Cox proportional hazards regression models.

ANOVA table An ANOVA table[15] illustrates which explanatory variables are related to the response variable [Grafen and Hails, 2006, Harrell, 2001, Harrell, 2012b]. Low p-values (conventionally $p < 0.05$) indicate that the null hypothesis for regression (H_0: a feature (x) adds no extra information to the output variable (y)) can be rejected, and quantify potential important differences towards the response variable.

[15] **R package**:
The following R functions outputs the ANOVA table of a linear model:
- anova: anova function has the linear model as an argument.

Coefficient table The coefficient (θ_i), or slope of the linear line, characterizes the linear relationship between a feature and the response variable [Grafen and Hails, 2006]. It is important to analyse the standard error[16] for each coefficient estimate, the corresponding p-value of the t-ratio is computed based on the standard error [Harrell, 2012b][17].

[16]
$$\text{t-ratio} = \frac{\text{estimate}}{\text{standard error}} \tag{2.1.2.7}$$

[17] **R package**:
The following R functions output the coefficient table of a linear model:
- coef: the coefficient table of the linear model.
- summary: the diagnostics output of the linear model.

Confidence intervals for model parameters The reliability of the parameters of a linear model can be analysed by their *confidence interval*. The confidence interval is calculated for a specific confidence level, e.g. 90%, 95%, and 99% levels are often used. Most statistical software packages provide specific functions for calculation of the confidence intervals[18].

[18] **R package**:
The following R function outputs the confidence interval:
- confint: returns an overview of the confidence interval of all the coefficients in the linear model.

$$\theta_i \pm t_{\frac{\alpha}{2},n-2} \times SE \tag{2.1.2.8}$$

The parameters of the linear model are often verified by *bootstrapping* (see section 2.1.5.1 in on page 82)[19] or another resampling method, e.g. cross-validation.

[19] **R package**:
The following R package can be used for bootstrapping of a linear model:
- boot [Canty and Ripley, 2012]: the boot function provides implementation for bootstrapping.

Confidence intervals for response variable If we want to predict the mean response of our linear model given a set of values for our features, $x^{(i)}$[20]:

$$E(y^{(j)}|x^{(j)}) = \theta_0 + \theta_1 x_1^{(j)} + \theta_2 x_2^{(j)} + \ldots + \theta_n x_n^{(j)} \tag{2.1.2.9}$$

$$\hat{y_{(j)}} \pm t_{\frac{\alpha}{2},n-k} \sqrt{MSE\left(x^{(j)^T}(X^TX)^{-1}x^{(j)}\right)} \tag{2.1.2.10}$$

The confidence interval for the response value concentrates on the *sampling error*. This sampling error represents the error that a linear models contains bacause it is not based on the complete population sample [Faraway, 2002a].

[20] **R package**:
The following R function outputs the confidence interval of the response variable:

- predict: returns an overview of the confidence interval of the response variable in the linear model (interval="confidence").

Prediction intervals for response variable The predicted value with a confidence interval and prediction interval are equal. The only difference is that the prediction interval is larger (extra *MSE* term in equations 2.1.2.10 and 2.1.2.12)[21]:

$$E(y^{(j)}|x^{(j)}) = \theta_0 + \theta_1 x_1^{(j)} + \theta_2 x_2^{(j)} + \ldots + \theta_n x_n^{(j)} \tag{2.1.2.11}$$

$$\hat{y_{(j)}} \pm t_{\frac{\alpha}{2},n-k} \sqrt{MSE\left(1 + x^{(j)^T}(X^TX)^{-1}x^{(j)}\right)} \tag{2.1.2.12}$$

The prediction interval for the response value concentrates on the *sampling error* and the variability around the predicted mean.

[21] **R package**:
The following R function outputs the prediction interval of the response variable:

- predict: returns an overview of the prediction interval of the response variable in the linear model (interval="prediction").

2.1.2.2 Model selection

If the constructed model lacks complexity and insufficiently fits the data, often called *underfitting* or *high bias*, the response variable (y) will not be well predicted. Alternatively, if the constructed model fits the data very well, and the error of the predicted response variable (y) is very low, there is a risk of *overfitting* or *high variance*. This overfitting means that new predictions of this linear model, based on another data set, can be still very poor.

Test model fit The variance is often used to test if the data sufficiently fits a model [Faraway, 2002b]. We compare the variance of the model ($\hat{\sigma}^2$) with the variance of the variables (σ^2).

If we compose a model that is not complex enough for our data, or has a wrong form, the estimated $\hat{\sigma}^2$ will be an overestimate. If our extracted model is too complex and over-fits the data, then $\hat{\sigma}^2$ will be underestimated [Faraway, 2002b].

The ratio of the true and estimated variance can be written as [Faraway, 2002b]:

$$\frac{\hat{\sigma}^2}{\sigma^2} \sim \frac{\chi^2_{n-p}}{n-p} \tag{2.1.2.13}$$

If there is a lack of fit:

$$\frac{(n-p)\,\hat{\sigma}^2}{\sigma^2} > \left(\chi^2_{n-p}\right)^{1-\alpha} \tag{2.1.2.14}$$

In order to improve the goodness of fit of a linear model, it is good to analysis two model properties [Harrell. et al., 1996]:

1. *Calibration*: in the case of the average predicted response variable of a regression model is equal to the actual average response variable, we say the regression model is good calibrated.
2. *Discrimination*: in the case of the variance of the predicted response variable of a regression model is positively correlated with the variance of the actual response variable.

2.1.2.3 Feature selection

The selection of the most significant features of the linear model is a fundamental step in model building. This model building follows the principle of Occam's Razor:

> Among several plausible explanations for a phenomenon, the simplest is best

This implies the simplest model is the one with the least number of explanatory variables. The more explanatory variables a model contains, the more degrees of freedom are not used efficiently [Faraway, 2002b]. There is not a golden rule to know the number of features (x_i) a linear model should contain. Harrell et. al. [Harrell. et al., 1996] suggest the following criterion based on the number of samples (m):

$$i < \frac{m}{10} \tag{2.1.2.15}$$

Akaike information criterion (AIC) Feature selection for a linear model is often performed by a stepwise Akaike information criterion. This criterion does not reject any statistical model. It is based on information theory and informs about how well data supports a model [Akaike, 1974]. The model with a minimum AIC is the best according this model selection value[22]. The stepwise procedure includes different explanatory variables at each step of the procedure; for each step the AIC is calculated.

$$AIC = -2\ln L_m + 2i \tag{2.1.2.16}$$

$$AIC = m\left(\ln(2\pi) + \ln\left(\frac{SSE}{m}\right) + 1\right) + 2(i+2) \tag{2.1.2.17}$$

Bayesian information criterion (BIC) Bayesian model averaging (BMA) can also be applied for feature selection in linear models. In order to perform model selection, it applies the Bayesian information criterion (BIC) [Schwartz, 1978][23]:

$$BIC = -2\ln L_m + \ln(n)i \tag{2.1.2.18}$$

$$BIC = m\left(\ln(2\pi) + \ln\left(\frac{SSE}{m}\right) + 1\right) + 2\ln(m)(i+2) \tag{2.1.2.19}$$

AIC and BIC are very similar, the main difference is that BIC penalizes over-parametrization based upon the sample size (m).

[22] **R package**:
The following R packages provides a stepwise AIC procedure for linear models:
- MASS: stepAIC function computes stepwise AIC procedure for general- and generalized linear models.
- stats4: AIC function.

[23] **R package**:
The following R packages provide Bayesian model averaging (BMA) for linear models:
- BMA [Raftery, 1995]: bicreg function computes BIC for general linear models.
- stats4: BIC function.

2.1.2.4 *Regularization*

Regularization is a method to avoid overfitting. Regularization adds an extra regularization term ($R_\lambda(\theta)$) that reduces the values of the parameters of the linear model (θ). For linear models, there are two main regularization methods: (1) *least absolute shrinkage and selection operator (LASSO)* [Tibshirani, 1995] and (2) *Tikhonov regularization* or *ridge regression* [Tikhonov, 1995]. Nowadays, there exist a huge variety of possible regularization terms [Fu, 1998]. Regression models with a regularization term are often called *penalized regression models*, which can be formulated as:

$$\hat{\theta} \arg\min_{\theta} \left[\sum_{j=1}^{m} \left(y^{(j)} - \sum_{i=1}^{n} x_i^{(j)} \theta_i \right) \right]^2 + R_\lambda(\theta) \tag{2.1.2.20}$$

$$R_\lambda(\theta) = \lambda \sum_{i=1}^{n} |\theta|^m \tag{2.1.2.21}$$

Table 2.13: Different regularization terms for linear models ($R_\lambda(\theta)$).

Regularization name	Regularization term
LASSO	$\lambda \sum_{i=1}^{n} \lvert\theta_i\rvert$
Ridge	$\lambda \sum_{i=1}^{n} {\theta_i}^2$
Adaptive LASSO	$\lambda \sum_{i=1}^{n} \frac{\theta_i}{\lvert\theta\rvert^\gamma}$

A complementary measure of overfitting is called *shrinkage* [Copas, 1983, Harrell. et al., 1996]. There are different applications for shrinkage. A first application is to measure overfitting (γ), often this measure can be used to correct the coefficient ($\gamma\theta X$) in a linear model. Shrinkage can be performed with bootstrapping, cross validation, and shrinkage heuristics. An example of a heuristic shrinkage estimator in a Cox regression model is given by Houwelingen and le Cessie [Copas, 1983, van Houwelingen J.C. and le Cessie S., 1990]:

$$\gamma = \frac{\chi^2_{\text{model}} - df - 1}{\chi^2_{\text{model}}} \tag{2.1.2.22}$$

This heuristic approach is very useful to have a quick estimate of the shrinkage in a model, and can be used as a remedie against overconfident too high or too low predictions [van Houwelingen J.C. and le Cessie S., 1990]. If we multiply all the predictor coefficients with the same shrinkage correction, we might still not have a very rigorous solution. Therefore, estimation of shrinkage factors should be based on bootstrapping and cross validation in combination with a penalized maximum likelihood estimate.

2.1.2.5 *Performance measures for linear models*

Coefficient of determination (R^2) One of the most often used performance measures for a linear model is the *coefficient of determination (R^2; often called* R squared*)*. It measures the amount of variance that is captured by the linear model [Grafen and Hails, 2006]. For a detailed description of the coefficient of determination, often called *R squared* (see section 2.1.5.5 on page 85).

2.1.2.6 *Interactions*

The model performance can be improved by including more complicated features that are based on the original features. One possible

strategy to find more informative features is to multiply two features, which are called *interactions*.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \beta_{12} x_1 x_2 + \beta_{23} x_2 x_3 + \ldots + \beta_{(k-1)k} x_{k-1} x_k + \epsilon \tag{2.1.2.23}$$

Alternatively, new features can be added into the linear model by applying transformation function on the original features (e.g., it can be transforming input data or or adding nonlinear terms) [Harrell, 2012b]. In table 2.14 you can find an overview of the most often used transformation functions.

Table 2.14: Transformation functions provided in R.

Function	Description
ns	natural spline
rcs	restricted cubic spline
bs	B-spline
log	logarithmic
poly	polynomial

2.1.2.7 *Generalized Linear Models (GeLM)*

The generalized linear models (GeLM) are an extension of the general linear models (GLM). GLM are described by equation 2.1.2.23 (see on page 53). A GeLM can be written as [Grafen and Hails, 2006][24]:

$$y = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n) + \epsilon \tag{2.1.2.24}$$

$$g^{-1}(y) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n + \epsilon \tag{2.1.2.25}$$

[24] **R package**:
The following R functions to support general linear models:

- glm: generalized linear model function that fits parameters θ.

The main difference between general- and generalized linear model is the introduction of the *canonical link function* ($g(y)$) and the inverse link function ($g^{-1}(y)$) [Harrell, 2012b]. These link functions are the main trick used to let the linear model think the response variable is still normally distributed. The link function maps predictor terms ($\Theta_i x_i$) and the response variable. The following table (2.15) provides an overview of the main link function used in the context of GeLM.

Table 2.15: The canonical- and inverse link function for Generalized linear models.

Distribution	Canonical link $\theta = g(y)$	Inverse link $y = g^{-1}(\theta)$
Poisson	$\log y$	$\exp y$
Binomial		
logit link	$\log\left(\frac{y}{1-y}\right)$	$\frac{\exp\theta}{1+\exp\theta}$
probit link	$\Phi^{-1}(y)$	$\Phi(y)$
cloglog link	$\log(-\log(1-\mu))$	$1-\exp(-\exp(\theta))$
Normal	y	θ
Gamma	$-\frac{1}{\mu}$	$-\frac{1}{\theta}$
Negative binomial	$\log(1-\mu)$	$1-\exp\theta$

During this PhD, a *logistic regression* model was constructed with generalized linear models. For logistic regression the link is typically the *logit link*[25] [Harrell, 2012b].

All the concepts that were explained for the general linear models, are also applicable for the generalized linear models. In the next sections, support vector machines (SVM) will be explained. They are one of the most powerful supervised machine learning algorithms.

[25] The logit function is also called the log odds: $\log\left(\frac{p}{1-p}\right)$

2.1.3 Support vector machines (SVM)

The advancements in *statistical learning theory (SLT)* during 1990s introduced a rigid framework for the generalization of machine learning algorithms [Boser et al., 1992]. A major practical outcome of this frameworks was the introduction of *large margin classifiers*. These classifiers learn the maximum margin of decision boundaries, meaning that not only a function will be used to classify, but also the maximum margins around this function will be considered. Today there are two main approaches for large margin classification: (1) *support vector machines* and (2) *boosting*. During this PhD, the support vector machines were mainly studied and applied.

Support vector machines (SVMs) are a set of learning algorithms which are called *Sparse Kernel Machines* [Bishop, 2006d]. After being introduced by Vapnik et. al. in 1992 [Boser et al., 1992], it became a popular approach to solve the problem of *classification*[26] in *supervised learning*. This new approach showed to be easier as the opaque neural networks [Press et al., 2007]. Very recently, neural networks start to regain popularity in the machine learning community. They are based on a paradigm to learn based on the propagation of information between neurons in the human brain. In the next paragraphs, a general formulation of the data set for classification will be provided. This will be fundamental for later formulations of support vector machines.

[26] Classification predict a binary output, e.g. a tumour is benign or malignant.

The data set that will be used for binary classification has specific properties. The data set will contain m samples and contain two parts: (1) n *explanatory variables*, also called *features* (x) and (2) the *response- or output variable* (y)[27]:

[27] $x_i^{(j)}$: is the value of the ith feature and the jth data sample.

$$\forall j \in \{1, 2, \ldots, m\} : d^{(j)} = \left\{x^{(j)}, y^{(j)}\right\} \tag{2.1.3.1}$$

$$\forall i \in \{1, 2, \ldots, n\} : x \in \mathbb{R}^{n \times m} \tag{2.1.3.2}$$

$$y \in \{-1, +1\} \tag{2.1.3.3}$$

Support Vector Machines provide a predicted value, not a probability [Boser et al., 1992]:

$$f(x) = h(x) + b \tag{2.1.3.4}$$

$$h(x) = \sum_i y_i \alpha_i k(x_i, x) \tag{2.1.3.5}$$

Our system under investigation has been tested with m observations. As an input for the SVM machine learning algorithms each observation consists of [Burges, 1998]:

1. *Explanatory variables or features (x)*: a feature vector exists in the continuous space $\mathbb{R}^n$, but can also contain binary features.
2. *Response variable or output variable (y)*: it would be 1 if the pattern is recognized and -1 if not.

The task of a SVM algorithm is to learn the mapping between input- and output space ($\chi : x_i \rightarrow Y : y$) [Burges, 1998]. The mapping

is called the hypothesis; it represents a function with adjustable parameters: $x \rightarrow f(x, \alpha)$. The adjustable parameters will be *trained* for the classifier.

Next, statistical learning theory and Lagrangian formulation are explained. They are required to understand the mathematics and logic of support vector machines. In the sections thereafter, statistical learning theory and Lagrangian formulation will be applied to derive different support vector formulations. Finally, feature selection strategies specific for support vector machine will be described.

2.1.3.1 Statistical learning theory

A statistical learning theory formulates a mathematical theory with four properties [Vapnik, 2010b, Vapnik, 2010a]:

1. *Consistency*: a learning algorithm must contain consistency conditions based on the *empirical risk minimization (ERM)*.
2. *Convergence*: quantification of the rate of convergence towards an optimal learning algorithm.
3. *Generalization*: a learning algorithm must contain conditions that guarantee performance on new data, i.e., avoid problems like overfitting, underfitting, etc.
4. *Algorithm*: a learning algorithm should be easy to implement and contain all previous properties.

Different hypotheses for the mapping between the input- and output space ($\mathcal{X} : x_i \rightarrow \mathcal{Y} : y$) can be compared with a *loss function* or *risk function*. This loss function specifies the difference between the estimated response variable and the response variable of the training set. There are numerous examples of loss functions, e.g., squared loss, absolute value loss, zero-one loss, log loss, etc.

As an example the absolute loss function can be written as:

$$l = \left| y^{(j)} - f(x^{(j)}, \alpha) \right| \tag{2.1.3.6}$$

The loss function (l), the hypothesis space ($\mathcal{H}$), and the probability error measure P_{XY} allow to formulate the *expected risk function* $R[h]$[28]:

$$R[h] = \int_{\mathcal{XY}} l(f(x, \alpha), y) dP_{XY} \tag{2.1.3.7}$$

For every hypothesis (h), there exist a corresponding *prediction error*[29] $R[h]$. The learning task corresponds to find the optimal hypothesis h^* characterized by finding the infimum[30] of the expected risk function [Vapnik, 2010a]:

$$h^* = \arg \inf_{h \in \mathcal{H}} R[h] \tag{2.1.3.8}$$

Since the underlying probability P_{XY} is unknown, the expected risk ($R[h]$) for an hypothesis (h) can not be computed. A possible

[28] In some literature the expected risk is also called actual risk [Burges, 1998].

[29] In some literature the prediction error is also generalization error.

[30] infimum is also called the greatest lower bound.

approach to infer the probability P_{XY} would be to estimate the conditional probability $P_{XY}(y|x)$ from the training set. This estimation can be hard when the quantity of features is high. The training samples required to estimate $P_{XY}(y|x)$ grows exponentially with the amount of features.

Statistical learning theory uses a different methodology to derive the optimal hypothesis h^*. Instead of sampling the probability distribution P_{XY}, the optimal hypothesis is derived from the training data set. The condition that is used to select the optimal hypothesis is called an *induction principle* [Vapnik, 2010a].

Empirical risk minimization induction principle The empirical risk function $R_{emp}[h]$ can be formulated based on the mean error on the predicted training samples. The loss function l, sample size m, and hypothesis space $\mathcal{H}$ allow us to formulate the empirical risk [Vapnik, 2010a]:

$$R_{emp}[h] = \frac{1}{m}\sum_{j=1}^{m} l(f(x^{(j)}), y^{(j)}) \tag{2.1.3.9}$$

The empirical risk induction principle is a very limited approach for finding the optimal hypothesis (h^*). This induction principle leads to an ill-posed mathematical problem, poor generalization conditions (no avoidance of under- and overfitting), etc. Therefore new induction principle will be introduced into the next paragraphs.

Induction principle based on regularization theory The regularization theory aids to formulate the regularized risk function based on the empirical risk, a term that penalizes high complexity of the hypothesis space $\mathcal{H}$ ($\Omega(h)$), and empirical training error (λ):

$$R_{reg}[h] = R_{emp}[h] + \lambda\Omega(h) \tag{2.1.3.10}$$

Structural risk minimization induction principle The complexity of the hypothesis space ($\mathcal{H}$) is a very important measure for obtaining the risk function for a machine learning algorithm. The hypothesis space are all combination of decision functions, e.g. binary classifier, rank classifier, regression, etc.. A potential way to measure the complexity of the hypothesis space is to derive all possible combinations of output assignments that the decision functions should predict correctly. Obviously, this is not evident to formulate theoretically. Therefore, we will introduce one of the most often used approximations: *Vapnik Chervonenkis entropy* and *Vapnik Chervonenkis dimension* [Vapnik, 2010a].

The *random VC entropy* of an hypothesis space is formulated as:

$$H_{VC} = \log_2 N^{\mathcal{H}}(x^{(1)}, x^{(2)}, \ldots, x^{(m)}) \tag{2.1.3.11}$$

The logarithm of the maximum of the *random VC entropy* is called

the *growth function*:

$$G_{\mathcal{H}}(m) = \log_2 \left(\max_{(x^{(1)},x^{(2)},\ldots,x^{(m)})} N^{\mathcal{H}}(x^{(1)}, x^{(2)}, \ldots, x^{(n)}) \right) \tag{2.1.3.12}$$

$$G_{\mathcal{H}}(m) = m \ln 2 \tag{2.1.3.13}$$

$$G_{\mathcal{H}}(m) \leqslant VC_{dim} \ln \frac{m}{VC_{dim}} + 1 \tag{2.1.3.14}$$

The Vapnik Chervonenkis dimension (VC dimension; VC_{dim}) are the largest number of points that can be predicted by hypothesis h of hypothesis space $\mathcal{H}$.

Risk bound In this case, the loss can only take values 0 or 1. We choose a η such that $0 \leqslant \eta \leqslant 1$. The following bound holds for losses with a probability of $1 - \eta$:

$$R(\alpha) \leqslant R_{emph}(\alpha) + \sqrt{\left(\frac{VC_{dim}\left(\log \frac{2l}{VC_{dim}} + 1\right) - \log \frac{\eta}{4}}{l} \right)} \tag{2.1.3.15}$$

VC_{dim} is a non-negative integer called the Vapnik Chervonenkis (VC) dimension.

VC confidence is the second term at the right side.

There are three major consequences of the risk bound:

1. Independent on $P(x, y)$: it assumes only that the training set and the test data are independent according to $P(x, y)$.
2. It is often not possible to compute $R(\alpha)$.
3. If VC_{dim} is known, it is possible to compute the left hand side.

VC dimension VC dimension is a property of a set of functions: $\{f(\alpha)\}$. We will only take into account the two-class pattern recognition case ($f(x, \alpha) \in \{-1, 1\}$). If m observed data samples can be labeled in 2^m possible ways and a set of parameters of $\{f(\alpha)\}$ can be found that retrieves the exact labels, we say that this set of data points is *shattered* by that set of functions.

The VC dimension of a set of functions $\{f(\alpha)\}$ is defined as the maximum number of data points that can be shattered by $\{f(\alpha)\}$. From equation 2.1.3.15 on page 57, the risk bound where VC dimension is represented by VC_{dim}, than there exists at least one set of VC_{dim} points that can be shattered, but not every set of VC_{dim} observed data point can be shattered.

Optimization theory Constructing a SVM machine learning model is based on *optimization theory*. In the following section primal- and dual optimization problem will be illustrated in the context of SVMs. Such an optimization problem is characterized by a quadratic objective function ($f(x)$) and linear constraints.

In practice, these optimization problems are solved with the *Lagrangian formulation*. The following paragraphs illustrate the mathematical computations required to solve the optimization problems for support vector machines.

2.1.3.2 *Lagrangian formulation*

In order to solve the SVM inequality and maximize the margin of the separating hyperplane (see equation 2.1.3.27 on page 61) *Lagrange multipliers* are used. There are two main advantages for applying Langrangian formulation to solve the constrained optimization problem: (1) the Lagrange multipliers can take constraints into account and (2) the training data will only appear in dot products between vectors.

This type of optimization problem retrieves the stationary points[31] of a function with several variables that are under one or more constraints [Bishop, 2006d].

[31] A stationary point is a point where the derivative of a function is zero.

$$\begin{aligned} &\text{Maximize :} && f(x_1, x_2) \\ &\text{Subject to :} && g(x_1, x_2) = 0 \end{aligned}$$

The optimization problem can be solved by finding a function of x_1 to express x_2 in the form of $x_2 = h(x_1)$. This can be substituted into $f(x_1, x_2)$ which leads to $f(x_1, h(x_1))$. The differentiation of $f(x_1, h(x_1))$ will provide the stationary value for x_1, which can provide you the corresponding stationary value of x_2 ($x_2 = h(x_1)$). The biggest drawback of this approach is the definition of $x_2 = h(x_1)$.

A more elegant solution of this problem is by the introduction of Lagrange multipliers.

Lagrange multipliers The Lagrange formula for solving a optimization problem:

$$L \equiv f(x, y) + \lambda \left(g(x, y) - c \right) \tag{2.1.3.16}$$

$f(x, y)$ function that needs to be maximized.

$g(x, y) = c$ constraint of this optimization problem.

λ are the Lagrange multipliers. They are the *stationary points* for the Lagrange function (λ can be positive or negative).

Geometrical interpretation A D-dimensional variable x with components $(x_1, \ldots, x_D)$ and a constraint equation ($g(x) = 0$) represents a D-1-dimensional surface in x-space illustrated by figure 2.9.

At any point of the constraint surface ($g(x) = 0$), the gradient of the constraint function ($\Delta g(x)$) will be orthogonal on the surface. This can be formulated by the application of a Taylor series on the point $x + \epsilon$:

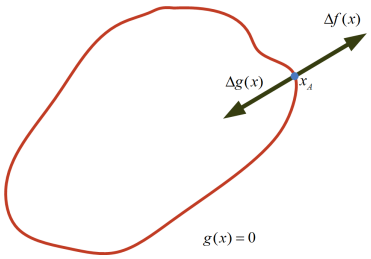


Figure 2.9: Geometrical interpretation of Lagrangian multiplier.

$$g(x + \epsilon) \simeq g(x) + \epsilon^T \Delta g(x) \tag{2.1.3.17}$$

Since x and $x + \epsilon$ are both on the constraint surface, we can write $g(x) = g(x + \epsilon)$ and $\epsilon^T \Delta g(x) \simeq 0$. In the limit:

$$\lim_{\|\epsilon\| \to 0} \epsilon^T \Delta g(x) = 0 \tag{2.1.3.18}$$

Hence ϵ is parallel to the constraint surface and the vector Δg is a normal to the surface.

The aim of this approach is to find the point on the surface of the constraint so that function $f(x)$ is maximized. Such a point has the property that the vector $\Delta f(x)$ is orthogonal to the constraint surface. Δf and Δg are parallel (or anti-parallel) vectors. There exists a parameter λ such that the following equation can be written:

$$\Delta f + \lambda \Delta g = 0 \tag{2.1.3.19}$$

The λ parameter is called the *Lagrangian multiplier*, they can be positive or negative. In analogy with equation 2.1.3.16, the Lagrangian formula can be written:

$$L(x, \lambda) \equiv f(x) + \lambda g(x) \tag{2.1.3.20}$$

The constrained stationary condition (see equation 2.1.3.19) can be derived because $\Delta_x L = 0$. The condition $\frac{\partial L}{\partial \lambda} = 0$ leads to the constraint condition $g(x) = 0$.

Finding the maxima of a function $f(x)$ subject to constraint $g(x) = 0$ can be done by solving formula 2.1.3.20. This allows us to compute the stationary points of $L(x, \lambda)$.

The problem of finding a maximum of a function can also occur with inequality constraints ($g(x) \geqslant 0$). Based on the constraint inequality, there are two possible solutions:

1. Constraint is *inactive*: the stationary point is situated in the region: $g(x) > 0$. The function $g(x)$ has no influence and the stationary condition is $\Delta f(x) = 0$. This corresponds to equation 2.1.3.20, but this time with $\lambda = 0$.

2. Constraint is *active*: the stationary point is situated on the boundary: $g(x) = 0$. This is equal to the equality constraint scenario ($\lambda \neq 0$). The sign of the Lagrangian multiplier will be crucial, since the function ($f(x)$) will be its gradient ($\Delta f(x)$); it is oriented away from the region $g(x) > 0$. So we can state that $\Delta f(x) = -\lambda \Delta g(x)$ for $\lambda > 0$.

In both cases, the product $\lambda g(x)$ is equal to zero. We can now state the following conditions for finding the maximum subject to the inequality constraints:

$$g(x) \geqslant 0 \tag{2.1.3.21}$$

$$\lambda \geqslant 0 \tag{2.1.3.22}$$

$$\lambda g(x) = 0 \tag{2.1.3.23}$$

These conditions are called the *Karush-Kuhn-Tucker (KKT) complementarity conditions*.

Primal and dual problems in quadratic programming

Primal problem A primal problem in quadratic programming can be written as [Press et al., 2007]:

$$\begin{array}{ll} \text{minimize:} & f(w) \\ \text{subject to:} & g_j(w) \leqslant 0 \\ & h_k(w) = 0 \end{array}$$

$f(w)$ quadratic function in w

$g_i(w) \leqslant 0$ inequality constraints in w

$h_k(w) = 0$ equality constraints in w

A Lagrangian that incorporates a quadratic form with all constraints can be written as:

$$L_p \equiv \frac{1}{2} f(w) + \sum_j \alpha_j g_j(w) + \sum_k \beta_k h_k(w) \tag{2.1.3.24}$$

Dual problem Every primal problem can be reformulated into a dual problem, which can be used as an alternative of solving the primal problem. The transformation from a primal to a dual problem starts with composing the subset of conditions for an extremum:

$$\frac{\partial L_p}{\partial w_i} = 0, \frac{\partial L_p}{\partial \beta_k} = 0 \tag{2.1.3.25}$$

This resulting equation will be used to substitute w and L_p by α and β, which leads to the dual for L_D:

$$\begin{array}{ll} \text{maximize:} & L_D \\ \text{subject to:} & \forall j : \alpha_i \geqslant 0 \end{array}$$

If $\hat{x}$ is the optimal solution of the primal problem, and $\hat{\alpha}$ and $\hat{\beta}$ the optimal solutions of the dual problem, we can write:

$$\begin{array}{ll} & f(\hat{w}) = L_{\hat{\alpha},\hat{\beta}} \\ \forall j & \hat{\alpha_j} g_j(\hat{w}) = 0 \end{array}$$

This last condition is called the *Karush-Kuhn-Tucker (KKT) complementarity condition*; it states that at least one $\hat{\alpha_j}$ and $g_j(\hat{w})$ must be zero for each j.

This Lagrangian theory will be used to formulate the optimization problem for support vector machines in different cases: linear separable case, linear non-separable case, and nonlinear case.

2.1.3.3 Support vector machines in the linear separable case

Support vector machines in the linear separable case are the most trackable to compute. They illustrate very well how optimization theory can be applied for deriving the separating hyperplane. First, we define a set of symbols that will be important for the following paragraphs.

training data $\{x_i, y_i\}, i = 1, \ldots, m, y_i \in \{-1, 1\}, x_i \in R^d$

separating hyperplane with highest margin $w.x + b = 0$: the equation of a straight line.

- w: normal to the hyperplane.
- $\frac{|b|}{\|w\|}$: orthogonal distance from the data point to the linear hyperplane. $\|w\|$ is the Euclidean norm of w.
- d_- and d_+: shortest distance from the separating hyperplane to the closest positive or negative data point.

$$\begin{aligned} x_i.w + b &\geqslant +1 \quad \text{for } y_i = +1; \\ x_i.w + b &\leqslant -1 \quad \text{for } y_i = -1; \end{aligned} \tag{2.1.3.26}$$

It can be combined into one set of inequalities:

$$y_i(x_i + b) - 1 \geqslant 0 \tag{2.1.3.27}$$

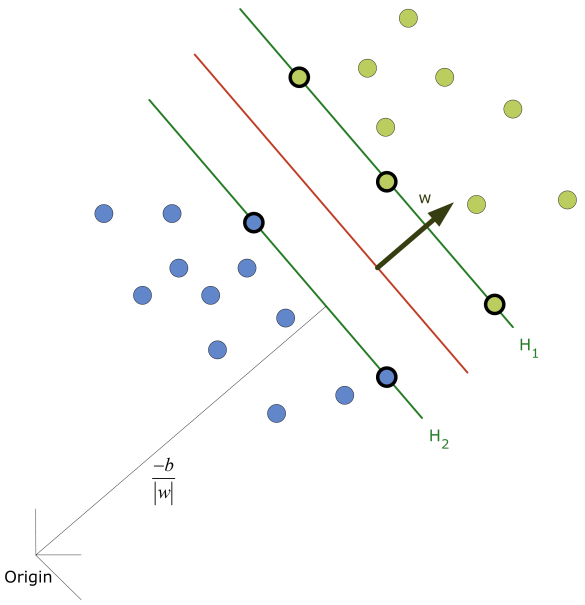


Figure 2.10: Support vector machine (SVM) in the linear separable case.

The data points with $f(x) = +/-1$ are those points that specify the maximum margin for the separating hyperplanes, they are called *support vectors* [Press et al., 2007]:

$$H_1 : x_i.w + b = 1 \tag{2.1.3.28}$$

$$H_2 : x_i.w + b = -1 \tag{2.1.3.29}$$

H_1 : orthogonal distance from the origin can be written as $\frac{|1-b|}{\|w\|}$

H_2 : orthogonal distance from the origin can be written as $\frac{|-1-b|}{\|w\|}$

$d_+ = d_- = \frac{1}{\|w\|}$

$\frac{2}{\|w\|}$: the margin between both hyperplanes.

H_1 and H_2 are parallel. The maximum margin will be computed by finding the maxima of $\|w\|^{-1}$, which is equivalent of finding the minima of $\|w^2\|$ [Bishop, 2006d]. We can formulate a optimization problem with a quadratic programme:

$$\begin{array}{ll} \text{minimize:} & \frac{1}{2}\|w\|^2 \\ \text{subject to:} & y_i\left(x_i.w+b\right) \end{array}$$

Lagrangian formulation for SVM in the linear separable case This optimization problem can be transformed in a Lagrangian formulation [Burges, 1998], and will illustrate how this optimization problem can be computed. The same procedures as in the previous paragraphs will be used. First, the primal formulation will be composed, which can be reformulated into a dual form. This dual form has often easier constraints and leads to a simpler computation of the optimization problem.

Primal Lagrangian formulation

$$L_P \equiv \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i y_i \left(x_i.w+b\right) + \sum_{i=1}^{m} \alpha_i \tag{2.1.3.30}$$

The subset of conditions for extrema:

$$\frac{\partial L_p}{\partial w_i} = w - \sum_i \alpha_i y_i x_i = 0 \Rightarrow \hat{w} = \sum_i \hat{\alpha_i} y_i x_i \tag{2.1.3.31}$$

$$\frac{\partial L_p}{\partial b} = \sum_i \alpha_i y_i = 0 \tag{2.1.3.32}$$

Karush-Kuhn-Tucker (KKT) complementarity condition These conditions are the general optionality conditions. They follow from *strong duality* and *complementarity*, and play a central role in the theory and practice of constrained optimization [Press et al., 2007].

For the primal problem, the KKT conditions can be stated [Burges, 1998]:

$$\frac{\partial L_p}{\partial w_v} = w_v - \sum_i \alpha_i y_i x_{iv} = 0; v = 1, \ldots, d \tag{2.1.3.33}$$

$$\frac{\partial L_p}{\partial b} = -\sum_i \alpha_i y_i = 0 \tag{2.1.3.34}$$

$$y_i(x_i.w+b) - 1 \geqslant 0; i = 1, \ldots, m \tag{2.1.3.35}$$

$$\forall i : \alpha_i \geqslant 0 \tag{2.1.3.36}$$

$$\forall i : \alpha_i(y_i(w.x_i+b)) = 0 \tag{2.1.3.37}$$

$$\tag{2.1.3.38}$$

Dual Lagrangian formulation The substitution of the extrema conditions into the primal Lagrangian formulation allow us to formulate

the dual Lagrangian formulation:

$$\begin{array}{ll} \text{Maximize:} & L_D = \sum_j \alpha_j - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i x_j \\ \text{Subject to:} & 0 \leqslant \alpha_i \\ & \sum_i \alpha_i y_i = 0 \end{array} \tag{2.1.3.39}$$

In order to solve the SVM problem, we need to solve the KKT conditions. The solution of this optimization problem is not expressed in w and b, but are reformulated in dual form. This dual form reformulates a solution for the classification function. This solution is reduced to the calculation of the Lagrangian multipliers.

Only the data samples of our training set that have Lagrangian multipliers not equal to zero ($\alpha_i \neq 0$) are required to solve the optimization problem. These data points are the so-called *support vectors*.

The definition of the dual optimization problem is not directly dependent on the training data set, but on their mutual inner products ($x_i x_j$).

2.1.3.4 *Support vector machines in the linear non-separable case*

In real world machine learning problems, it seldom occurs that the data can be separated linearly. In the case of linear non-separable data, an alternative formulation is needed to classify the data. It is necessary to weaken the constraints of the linear separable SVM (see equation 2.1.3.40 on page 63). Therefore, *slack variables* ($\xi_i; i = 1, \ldots, m$) can be introduced into the constraints:

$$\begin{array}{ll} x_i.w + b \geqslant +1 - \xi_i & \text{for } y_i = +1; \\ x_i.w + b \leqslant -1 + \xi_i & \text{for } y_i = -1; \\ \forall i : \xi_i \geqslant 0 & \end{array} \tag{2.1.3.40}$$

$\sum_i \xi_i$: is the upper bound on the number of training errors.

The natural way to assign the extra cost of errors is to change the objective function to be minimized from $\frac{\|w\|^2}{2}$ to $\frac{\|w\|^2}{2} + C \left(\sum_i \xi_i\right)^k$.

C : end user-defined parameter to configure the penalty to errors. A higher C corresponds to a higher penalty to errors.

This is a convex programming problem for any integer k. The 2-norm soft margin SVM ($k = 2$) and 2-norm soft margin SVM ($k = 1$) are quadratic programming problems and can be computed in polynomial time [Press et al., 2007].

$$\begin{array}{ll} \text{minimize:} & \frac{\|w\|^2}{2} + C \left(\sum_i^m \xi_i\right) \\ \text{subject to:} & x_i.w + b \geqslant 1 - \xi_n \\ & \xi_i \geqslant 0 \end{array}$$

Slack variables The slack variables ($\xi_i; i = 1, \ldots, l$) have specific values corresponding the data training point [Vapnik, 2010a]:

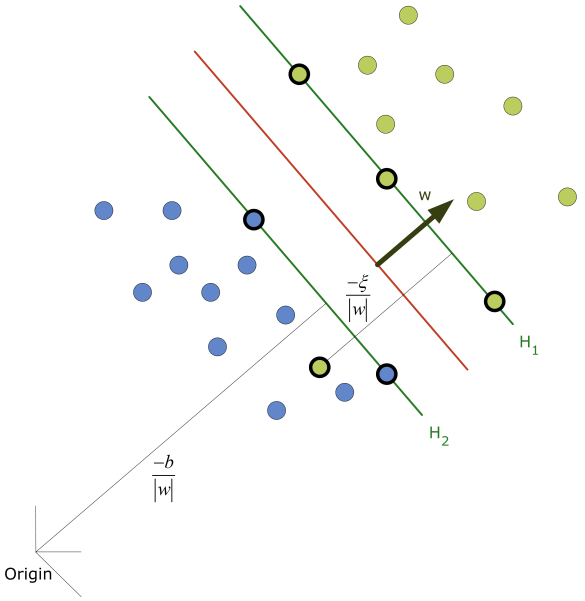


Figure 2.11: Support vector machine (SVM) in the non-separable case.

$\xi_l = 0$ data points are on the correct margin or in the correct side for classification.

$0 < \xi_l \leqslant 1$ data points lie inside the margin, but on the correct side of the decision boundary.

$\xi_l > 1$ data points lie on the wrong side of the decision boundary and are misclassified.

Primal Lagrangian formulation

$$L_P = \frac{1}{2}\|w\|^2 + C\sum_i \xi_i - \sum_i \alpha_i \left[y_i(x_i.w + b) - 1 + \xi_i\right] - \sum_i \mu_i \xi_i \tag{2.1.3.41}$$

Karush-Kuhn-Tucker (KKT) complementarity conditions The new Lagrange multipliers (μ_i) are introduced to enforce positivity of the ξ_i. The corresponding KKT conditions are given by:

$$\frac{\partial L_P}{\partial w_v} = w_v - \sum_i \alpha_i y_i x_{iv} = 0 \tag{2.1.3.42}$$

$$\frac{\partial L_P}{\partial b} = -\sum_i \alpha_i y_i = 0 \tag{2.1.3.43}$$

$$\frac{\partial L_P}{\partial \xi_i} = C - \alpha_i - \mu_i = 0 \tag{2.1.3.44}$$

$$y_i(x_i.w + b) - 1 + \xi_i \geqslant 0 \tag{2.1.3.45}$$

$$\xi_i \geqslant 0 \tag{2.1.3.46}$$

$$\alpha_i \geqslant 0 \tag{2.1.3.47}$$

$$\mu_i \geqslant 0 \tag{2.1.3.48}$$

$$\alpha_i \left\{y_i(x_i.w + b) - 1 + \xi_i\right\} = 0 \tag{2.1.3.49}$$

$$\mu_i \xi_i = 0 \tag{2.1.3.50}$$

Dual Lagrangian formulation

$$\begin{aligned} &\text{Maximize:} && L_D \equiv \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i.x_j \\ &\text{Subject to:} && 0 \leqslant \alpha_i \leqslant C \\ & && \sum_i \alpha_i y_i = 0 \end{aligned} \tag{2.1.3.51}$$

The dual form for the linear non-separable- and the separable case have similar formulations. Only the Lagrangian multipliers (α_i) are constrained by user-defined regularization factor C in the non-separable case. The C parameter is trade-off between the margin width and training error. The above formulation is referred as *C-SVM formulation* [Burges, 1998, Boser et al., 1992].

2.1.3.5 Support vector machines in non-linear case

Many real world applications for SVMs are demanding for a non-linear decision function. The dual optimization problem of SVMs in the linear separable- and non-separable case depend on the mutual inner products. This property is used together with the *reproducing kernel Hilbert space (RKHS) theory* to introduce non-linearity for SVM problems [Vapnik, 2010a].

In order to make a non-linear SVM, we need to map our training data into the *Euclidean space* ($\mathcal{H}$). Let's call this mapping Φ:

$$\Phi : \mathbb{R}^d \rightarrow \mathcal{H} \tag{2.1.3.52}$$

This mapping is performed on the mutual inner product of the dual Lagrangian formulation. The *inner product kernel trick* allows to higher the dimensional space, but the computational complexity does not increase drastically and the curse of dimensionality is avoided [M. A. Aizerman and Rozonoér, 1964, Scholköpf, 2001]. The dot products $x_i.x_j$ in $\mathcal{H}$ is transformed after the mapping: $\Phi(x_i).\Phi(x_i)$. The *kernel function* can be formulated as:

$$K(x_i.x_j) = \Phi(x_i).\Phi(x_i) \tag{2.1.3.53}$$

When feature vectors are mapped from a lower-dimensional space to a higher-dimensional embedding space, non-linear separation surfaces can become well approximated by linear surfaces. In practice, very high dimensional embedded spaces are used. They enter the SVM calculation only implicitly, through the *kernel trick* [Press et al., 2007].

The dual Lagrangian for SVM in the non-linear case can be formulated as:

$$\begin{aligned} &\text{Maximize:} && L_D \equiv \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j K\left(x_i, x_j\right) \\ &\text{Subject to:} && 0 \leqslant \alpha_i \leqslant C \\ & && \sum_i \alpha_i y_i = 0 \end{aligned} \tag{2.1.3.54}$$

Reproducing kernel functions There is a set of kernel functions that are applied for SVM learning:

Radial basis function (RBF) kernel

$$K(x, y) = \exp\left(-\gamma \|x - y\|^2\right) \tag{2.1.3.55}$$

Polynomial kernel

$$K(x, y) = (\langle x, y \rangle + 1)^p \tag{2.1.3.56}$$

Sigmoidal kernel

$$K(x, y) = \tanh\left(\langle x, y \rangle - \theta\right) \tag{2.1.3.57}$$

ν-SVM formulation Schölkopf et al. [Scholköpf et al., 2000] proposed an alternative for the C-SVM formulation explained in the previous sections: the ν-SVM formulation. Its main contribution is to explicitly formulate the margin information for the SVM [Chen et al., 2005].

$$H_1 : x_i.w + b = \rho \tag{2.1.3.58}$$
$$H_2 : x_i.w + b = -\rho \tag{2.1.3.59}$$

The loss function for the ν-SVM formulation is written as:

$$L_\nu(x, y)) = \begin{cases} 0 & \text{if } |y - f(x)| \geqslant \rho \\ \rho - yf(x) & \text{otherwise} \end{cases} \tag{2.1.3.60}$$

The optimization problem for the soft margin classifier is formulated as [Press et al., 2007]:

$$\begin{array}{ll} \text{minimize:} & \frac{\|w\|^2}{2} - \nu\rho + \sum_i^m \xi_i \\ \text{subject to:} & x_i.w + b \geqslant \rho - \xi_i \\ & \xi_i \geqslant 0 \\ & \rho > 0 \end{array}$$

Primal Lagrangian formulation

$$\begin{array}{ll} \text{minimize:} & L_p = \frac{\|w\|^2}{2} - \nu\rho + \frac{1}{m}\sum_i^m \xi_i - \gamma\rho - \sum_i^m \alpha_i \left[y_i(x_i.w + b) - \rho + \xi_i\right] - \sum_i^m \mu_i\xi_i \\ \text{subject to:} & x_i.w + b \geqslant \rho - \xi_i \\ & \xi_i \geqslant 0 \\ & \rho > 0 \end{array}$$

Karush-Kuhn-Tucker (KKT) complementarity conditions

$$\frac{\partial L_P}{\partial w_v} = w_v - \sum_i^m \alpha_i y_i x_{iv} = 0 \tag{2.1.3.61}$$

$$\frac{\partial L_P}{\partial b} = -\sum_i^m \alpha_i y_i = 0 \tag{2.1.3.62}$$

$$\frac{\partial L_P}{\partial \xi_i} = -\rho - \nu - \alpha_i = 0 \tag{2.1.3.63}$$

$$y_i(x_i.w + b) - \rho + \xi_i \geqslant 0 \tag{2.1.3.64}$$

$$\xi_i \geqslant 0 \tag{2.1.3.65}$$

$$\alpha_i \geqslant 0 \tag{2.1.3.66}$$

$$\mu_i \geqslant 0 \tag{2.1.3.67}$$

$$\alpha_i \left[y_i(x_i.w + b) - \rho + \xi_i\right] = 0 \tag{2.1.3.68}$$

$$\mu_i \xi_i = 0 \tag{2.1.3.69}$$

$$\rho\gamma = 0 \tag{2.1.3.70}$$

Dual Lagrangian formulation The dual Lagrangian for νSVM formulation in the non-linear case can be formulated as:

$$\begin{array}{ll} \text{Maximize:} & L_D \equiv \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j K\left(x_i, x_j\right) \\ \text{Subject to:} & 0 \leqslant \alpha_i \leqslant \frac{1}{m} \\ & \sum_i \alpha_i y_i = 0 \\ & \sum_i \alpha_i \geqslant \nu \end{array} \tag{2.1.3.71}$$

In the following sections, the SVM learning theory for soft margin classifiers will be reformulated for regression problems.

2.1.3.6 *SVM for regression*

In SVM regression, the input x_i is mapped into a n-dimensional feature space using a fixed (non-linear) mapping. In this feature space a linear model is constructed. The linear model can be written as:

$$f(x, \omega) = \sum_{j=1}^{n} \omega_j g_j(x) + b \tag{2.1.3.72}$$

$g_i(x)$ set of non-linear transformations

b bias term

Regression estimates are computed by minimization of the empirical risk on the training data. SVM regression uses a new type of loss function, called *ϵ-insensitive loss function* introduced by Vapnik [Vapnik, 2010a].

$$L_\epsilon(y, f(x, \omega)) = \begin{cases} 0 & \text{if } |y - f(x, \omega)| \leqslant \epsilon \\ |y - f(x, \omega)| - \epsilon & \text{otherwise} \end{cases} \tag{2.1.3.73}$$

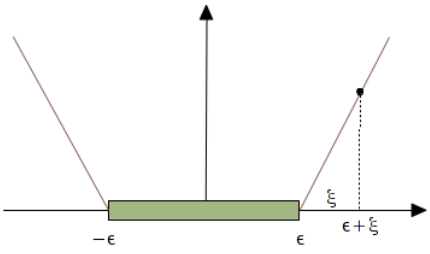


Figure 2.12: The soft margin loss for SVM regression [Smola and Schölkopf, 2004].

The empirical risk (R_{emp}) for SVM regression is formulated as:

$$R_{emp}(\omega) = \frac{1}{n}\sum_{i=1}^{n} L_\epsilon(y_i, f(x_i, \omega)) \tag{2.1.3.74}$$

The purpose of ϵ-SV regression is to find a function $f(x)$ with the maximum deviation (ϵ) for the predicted output from all the samples of the training data. An error is only important if it becomes bigger as the deviation ϵ

training data : $\mathbb{D} = \{x_i, y_i\} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\} \subset \chi \times \mathbb{R}$

space of input features : χ

Analogously with the SVM margin classifier formulations, we start from a linear the regression function ($f(x)$):

$$w \in \chi, b \in \mathbb{R} : f(x) = \langle w, x\rangle + b \tag{2.1.3.75}$$

$\langle .,. \rangle$ denotes a dot product in χ. This convex optimization problem can be written as:

$$\begin{array}{ll} \text{minimize :} & \frac{1}{2}\|w\|^2 \\ \text{subject to :} & y_i - \langle w, x_i\rangle - b \leqslant \epsilon \\ & \langle w, x_i\rangle + b - y_i \leqslant \epsilon \end{array} \tag{2.1.3.76}$$

Necessary in order to deal with the non-separable data, we introduce a soft margin loss function with *slack variables* ($\xi_i; i = 1, \ldots, m$; see section 2.1.3.4 on page 63). The slack variables weaken the constraints of the optimization problem (primal problem):

$$\begin{array}{ll} \text{minimize :} & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} \xi_i + \xi_i^* \\ \text{subject to :} & y_i - \langle w, x_i\rangle - b \leqslant \epsilon + \xi_i \\ & \langle w, x_i\rangle + b - y_i \leqslant \epsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geqslant 0 \end{array} \tag{2.1.3.77}$$

The regularization parameter (C) determines the trade-off between flatness of f and the amount of tolerated deviations (ϵ). This tolerated deviation is formulated by the *ϵ-insensitive loss function*. This convex optimization problem will be solved with the standard dualization method utilizing Lagrange multiplier as described in previous sections.

Lagrangian formulation for SVM regression

Primal Lagrangian formulation

$$L_P \equiv \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i + \xi i^* - \sum_{i=1}^{m} \eta_i\xi_i + \eta_i^*\xi_i^* - \sum_{i=1}^{m} \alpha_i\left(\epsilon + \xi_i - y_i + \langle w, x_i\rangle + b\right) - \sum_{i=1}^{m} \alpha_i^*\left(\epsilon + \xi_i + y_i - \langle w, x_i\rangle - b\right) \tag{2.1.3.78}$$

The primal Lagrangian formulation contains Lagrangian multipliers: α_i, α_i^*, η_i and η_i^*.

Karush-Kuhn-Tucker (KKT) complementarity conditions

$$\frac{\partial L_P}{\partial w} = w - \sum_i \left(\alpha_i^* - \alpha_i\right) x_i = 0 \tag{2.1.3.79}$$

$$\frac{\partial L_P}{\partial b} = -\sum_i \alpha_i^* - \alpha_i = 0 \tag{2.1.3.80}$$

$$\frac{\partial L_P}{\partial \xi_i} = C - \alpha_i - \eta_i = 0 \tag{2.1.3.81}$$

$$\frac{\partial L_P}{\partial \xi_i^*} = C - \alpha_i^* - \eta_i^* = 0 \tag{2.1.3.82}$$

$$\frac{\partial L_P}{\partial \eta_i} = \sum_{i=1}^{m} \xi_i = 0 \tag{2.1.3.83}$$

$$\frac{\partial L_P}{\partial \eta_i^*} = \sum_{i=1}^{m} \xi_i^* = 0 \tag{2.1.3.84}$$

The dual variables, η_i and η_i^*, are eliminated through the partial derivation($\frac{\partial L_P}{\partial \xi_i^{(*)}}$) and can be reformulated as:

$$\eta_i^{(*)} = C - \alpha_i^{(*)} \tag{2.1.3.85}$$

Several conclusions can be drawn from the KKT conditions:

1. Only samples (x_i, y_i) with corresponding $\alpha_i^{(*)} = C$ are situated out the ϵ-insensitive tube.

2. $\alpha_i \alpha_i^* = 0$, there is no set of dual variables α_i, α_i^* which are simultaneously nonzero.

From these conclusions, the following can be formulated:

$$\begin{array}{lll} \epsilon - y_i + \langle w, x_i \rangle + b \geqslant 0 & \xi_1 = 0 & \text{if } \alpha_i < C \\ \epsilon - y_i + \langle w, x_i \rangle + b \leqslant 0 & & \text{if } \alpha_i > 0 \end{array} \tag{2.1.3.86}$$

Dual Lagrangian formulation

$$\begin{array}{ll} \text{maximize} & \frac{1}{2} \sum_{i,j=1}^{l} \left(\alpha_i - \alpha_i^*\right)\left(\alpha_j - \alpha_j^*\right) \langle x_i, x_j \rangle - \epsilon \sum_{i=1}^{l} \alpha_i + \alpha_i^* + \sum_{i=1}^{l} \alpha_i - \alpha_i^* \\ \text{subject to} & \sum_{i=1}^{l} \alpha_i - \alpha_i^* = 0 \\ & \alpha_i, \alpha_i^* \in [0, C] \end{array} \tag{2.1.3.87}$$

The partial derivative $\frac{\partial L_P}{\partial w}$ can be reformulates as:

$$w = \sum_{i=1}^{l} \left(\alpha_i -' alpha_i^*\right) x_i \tag{2.1.3.88}$$

$$f(x) = \sum_{i=1}^{l} \left(\alpha_i - \alpha_i^*\right) \langle x_i, x \rangle + b \tag{2.1.3.89}$$

The function $f(x)$ is only dependent on the mutual inner product of samples in the data set, which again allows us to reformulate the linear SVR into a non-linear SVR [Basak et al., 2007].

The parameter b can be derived from the KKT conditions:

$$\alpha_i \left(\epsilon + \xi_i - y_i + \langle w, x_i \rangle + b\right) = 0 \tag{2.1.3.90}$$

$$\alpha_i^* \left(\epsilon + \xi_i^* + y_i - \langle w, x_i \rangle - b\right) = 0 \tag{2.1.3.91}$$

$$\left(C - \alpha_i\right) \xi_i = 0 \tag{2.1.3.92}$$

$$\left(C - \alpha_i^*\right) \xi_i^* = 0 \tag{2.1.3.93}$$

2.1.3.7 *Nonlinear SVM regression*

In order to perform nonlinear SVM regression, a mapping is required in analogy with SVM formulations for classification (see section 2.1.3.5 on page 65). This can be achieved by preprocessing the training patterns into a feature space $\mathbb{F}$:

$$\Phi : \xi \rightarrow \mathbb{F} \tag{2.1.3.94}$$

After this preprocessing is executed, the standard SV algorithm can be applied. One needs to be careful with the application, because this can become computationally infeasible. In order to overcome this limitation, kernels can bring a cheaper way of solving the problem.

Mapping with kernels The SV algorithm (see equation 2.1.3.87 on page 69) only depends dot products between patterns x_i. Therefore if we can state $k(x, x^{'}) = \langle \Phi(x), \Phi(x^{'}) \rangle$, than we do not need to define Φ explicitly. The SV optimization problem can be restated:

$$\begin{array}{ll} \text{maximize} & -\frac{1}{2} \sum_{i,j=1}^{l} \left(\alpha_i - \alpha_i^*\right)\left(\alpha_j - \alpha_j^*\right) k(x_i, x_j) - \epsilon \sum_{i=1}^{l} \left(\alpha_i + \alpha_i^*\right) + \sum_{i=1}^{l} y_i \left(\alpha_i - \alpha_i^*\right) \\ \text{subject to} & \sum_{i=1}^{l} \left(\alpha_i - \alpha_i^*\right) = 0 \\ & \alpha_i, \alpha_i^* \in [0, C] \end{array} \tag{2.1.3.95}$$

We can write w and $f(x)$ as:

$$w = \sum_{i=1}^{l} \left(\alpha_i - \alpha_i^*\right) \Phi(x_i) \tag{2.1.3.96}$$

$$f(x) = \sum_{i=1}^{l} \left(\alpha_i - \alpha_i^*\right) k(x_i, x_2) + b \tag{2.1.3.97}$$

2.1.3.8 *Other SVM formulations*

There are still SVM formulations that will not be handled within the course of this text. For the soft margin classifiers, there exist a μ-SVM formulation [Crisp and Burges, 1999] that will not be discussed. Also, the ν-SVM formulation for regression [Scholköpf et al., 2000] will not be discussed here[32,33].

[32] **SVM packages**:
The following packages supports different SVM formulations for classification and regression:

- libsvm [Chang and Lin, 2011].
- tinysvm [Kudoh, 2000].

[33] **R package**:
The following R package supports different SVM formulations for classification and regression:

- e1071 [Meyer et al., 2012]: R interface to the libsvm package.

2.1.3.9 *Feature selection for SVM regression*

Feature selection is a technique of selecting optimal feature set among original features set by removing irrelevant or redundant

features. The major advantages of feature selection are, e.g. increase systems interpretability, improve generalization performance, minimize the overfitting for some learning algorithms, etc.

There are two main types of feature selection in SVM: (1) *filter methods*: independent of the underlying machine learning algorithms and (2) *wrapper methods*: dependent of the underlying machine learning algorithms. The wrapper method is preferable in many applications, but can lead to high computational load.

Originally, feature selection was meant for the classification problem. For the regression problem, feature selection can suffer from important ordinal information loss.

F score The F score measures the discrimination between two sets of real numbers. The higher the F score, the easier it is to discriminate between the positive and negative instances of the training set [Chen and Lin, 2006].

$$F(i) \equiv \frac{\left(\bar{x}_i^+ - \bar{x}_i\right)^2 + \left(\bar{x}_i^- - \bar{x}_i\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}\left(x_{k,i}^+ - \bar{x}_i^+\right)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}\left(x_{k,i}^- - \bar{x}_i^-\right)^2} \tag{2.1.3.98}$$

number of positive and negative instances of the training set n_- and n_+.

average value of the complete training set, the negative and positive instances of the feature i $\bar{x}_i$, $\bar{x}_i^-$ and $\bar{x}_i^+$.

feature i of the kth negative/positive instance $x_{k,i}^-$ and $x_{k,i}^+$.

The numerator indicates the discrimination between positive and negative sets, and the denominator indicates the discrimination within each of the two sets.

After the SVM formulations, survival analysis will be described.

2.1.4 Survival analysis

Survival analysis is applied to describe and quantify time-to-event data [Stevenson, 2009]. This group of statistical approaches focuses on the distribution of survival time (T). It has been successfully used for a variety of purposes, e.g. overall survival (OS) in a clinical trail [Collett, 2004b], duration analysis in sociology and economics [van den Berg, 2001], reliability analysis in engineering [Henley and Kumamoto, 1999], etc.

First, we will introduce the survival analysis related terminology. Followed by the computational methodologies used for survival analysis, i.e. non-parametric- and semi-parametric approaches. Finally, feature selection techniques and performance measures for survival analysis will be discussed.

2.1.4.1 Terminology

Survival analysis and its terminology is derived from its prototypical event of death. There are two fundamental functions in a

survival analysis: (1) *survival function ($S(t)$)* and (2) *hazard function ($h(t)$)* [Collett, 2004b].

Survival function ($S(t)$) The time-to-event ($t > 0$) can be seen as a sample of a variable T with an underlying probability density function ($f(t)$). The distribution function of T can be formulated as:

$$F(t) = Pr(T < t) \tag{2.1.4.1}$$
$$f(t) = \frac{F(t)}{dt} \tag{2.1.4.2}$$
$$F(t) = \int_0^t f(u)du \tag{2.1.4.3}$$

The distribution function of T represents the probability that the survival time is less than some time value t. The proportion of occurrences that the event has happened as a function of t is called the *failure function (cumulative distribution function)* ($F(t)$).

Survival function ($S(t)$) The survival function ($S(t)$) is defined as probability that the survival time is greater than or equal to t. The survival function is the complement of the cumulative distribution function ($F(t)$):

$$S(t) = Pr(T \geqslant t) = 1 - F(t) \tag{2.1.4.4}$$

The area under the curve (AUC) to the right of t is proportional to the probability the event of interest has not occurred. The survival curve is determined from the instantaneous failure rate curve. The survival function represents the probability that an individual survives for time t.

$$S(t) = exp(-H(t)) \tag{2.1.4.5}$$
$$S(t) = S_0(t)^{e^{\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{k1}}} \tag{2.1.4.6}$$
$$S_0(t) = \exp\left(-H_0(t)\right) \tag{2.1.4.7}$$

Instantaneous hazard ($h(t)$) Another representation of the distribution of survival times is the *hazard function*, also often called *risk*. It is the conditional probability that a random individual will die at time $t + \delta t$ given that the individual has survived until time t. Instantaneous hazard ($h(t)$), or also called conditional failure rate, force of mortality, risk of death, etc.

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr\left(t \leqslant T < t + \Delta t | T \geqslant t\right) \geqslant t}{\Delta t} \tag{2.1.4.8}$$
$$h(t) = \frac{f(t)}{1 - F(t)} \tag{2.1.4.9}$$
$$h(t) = \frac{f(t)}{S(t)} \tag{2.1.4.10}$$
$$h(t) = -\frac{d}{dt}(\ln S(t)) \tag{2.1.4.11}$$

Cumulative hazard ($H(t)$), also called integrated hazard, equal the area under the instantaneous hazard until time t. The relationship between cumulative hazard and survival is as follows:

$$H(t) = \int_0^t h(u)du \tag{2.1.4.12}$$

$$H(t) = -\ln S(t) \tag{2.1.4.13}$$

$$S(t) = \exp\left(-H(t)\right) \tag{2.1.4.14}$$

The hazard function, or sometimes the logarithmic hazard function, can be fitted into a standard distribution, e.g., if the hazard function is constant ($h(t) = \alpha$) corresponds with a exponential distribution of times (see table 2.16).

Table 2.16: Joint probability distribution table of our small example.

Hazard function	Density function
$h(\alpha)$	Exponential distribution
$\log h(t) = \alpha + pt$	Gompertz distribution
$\log h(t) = \alpha + p \log t$	Weibull distribution

Censoring Survival analysis implies a follow-up period in order to collect the data. The start point and end point of this follow-up period could lead to incomplete information, this is called *censoring*. Censoring can occur when the event did not occur in the course of the study, when the person was lost during follow-up, etc. The data in a survival analysis is of the form:

$$d^{(j)} = \left\{T^{(j)}, \delta^{(j)}, x_i^{(j)}\right\} \tag{2.1.4.15}$$

Survival analysis data is a collection of the *failure time* or *censoring time* ($T^{(j)}$), the *censoring indicator* ($\delta^{(j)}$; $\delta = 0$ *censored and* $\delta = 1$ *failure time)*, and the *set of features* $x_i^{(j)} = \left\{x_1^{(j)}, x_2^{(j)}, \ldots, x_n^{(j)}\right\}$.

Censoring appears in two basic types [Kalbfleisch and Prentice, 2002, Kleinbaum and Klein, 2005]: (1) *type I censoring* and (2) *type II censoring*. Type I censoring, also called *time censoring* is not event-driven, meaning that within the pre-defined time frame of observation the quantity of failures is random. New items can enter the study at random, therefore this type of censoring appears often in medical research. Type II censoring, or *order statistic censoring*, is event driven. Hereby, a survival analysis will continue until a predefined quantity of failures occurred, e.g. perform the survival analysis for 5, 10, and 15 failures under different temperature 20 °C, 25 °C, and 30 °C.

During this PhD we modelled data with type I censoring (see chapter 3 on page 3). There are three forms of type I censoring [van den Berg, 2001, Collett, 2004b]:

1. *Right censoring*: if the event of interest occurs *after* the recorded follow-up period. For example, a patient is still alive after the period of observation.

2. *Left censoring*: if the event of interest occurs *before* the recorded follow-up period. For example, when the initial risk at the case is unknown.

3. *Interval censoring*: when left and right censoring occurs together.

Kalbfleisch and Pentice [Kalbfleisch and Prentice, 2002] introduced two important characteristics of censored data in statistical modelling, censoring is (1) *independent* and (2) *non-informative*.

The censoring is independent, if the hazard ($h(t)$) at any given time (t) would be the same with or without censoring. Whereas, non-informative means that the model parameters (θ) are not depending on censoring time ($T^{(j)}(\delta^{(j)} = 0)$) [Kalbfleisch and Prentice, 2002].

2.1.4.2 *Non-parametric approaches*

After collecting the time-to-event data, we would like to construct a model for that data. The first approach is to visualize the data by the survival curve. In this way we can identify the appropriate distribution for the data.

There are three non-parametric techniques for survival analysis: (1) *Life table method*, (2) *Kaplan-Meier*, and (3) *Nelson-Aalen*. In the next sections we will illustrate Kaplan-Meier and Nelson-Aalen. The life table will not be explained during the course of this text.

Kaplan-Meier method The Kaplan-Meier survival estimator is based upon individual survival times and assumes that the censoring is independent of survival time. It can be formulated as, for $0 \leqslant t \leqslant t^{+}$:

$$\hat{S}(t) = \prod_{j;t^{(j)} \leqslant t} \frac{r^{(j)} - d^{(j)}}{r^{(j)}} \tag{2.1.4.16}$$

The total set of collected failure times ($t^{(j)}$) in the course of a study is ordered. Each failure time has a corresponding amount of failures ($d^{(j)}$) and the number of individuals that are at risk ($r^{(j)}$) [Collett, 2004b]. This set of measures allow us to plot survival curves and make estimates on the survival probability at a given time. Often these curves are plotted for different samples of the data, e.g. different treatment regimen for a patient, different stages of a tumour, etc.

A nonparametric hypothesis test, that is often used complementary to Kaplan-Meier curves is the *log-rank* test [Kleinbaum and Klein, 2005]. This test will provide a p-value that indicates how different two samples are, e.g. a clinical trial can be tested if treatment regimen 1 is better as treatment regimen 2[34].

[34] **R package**:
The survival R packages provides Kaplan-Meier estimate and log-rank test:
- survfit: Kaplan-Meier and Flemington-Harrington estimate.
- survdiff: log-rank test.

Nelson-Aalen method The Nelson-Aalen method estimates cumulative hazard at time t [Collett, 2004b, Stevenson, 2009]:

$$\hat{H}(t) = \sum_{j;t^{(j)} \leqslant t} \frac{d^{(j)}}{r^{(j)}} \tag{2.1.4.17}$$

The Flemington-Harrington estimate of survival can be calculated using the Nelson-Aalen estimator (see equation 2.1.4.14 on page 73) [Collett, 2004b].

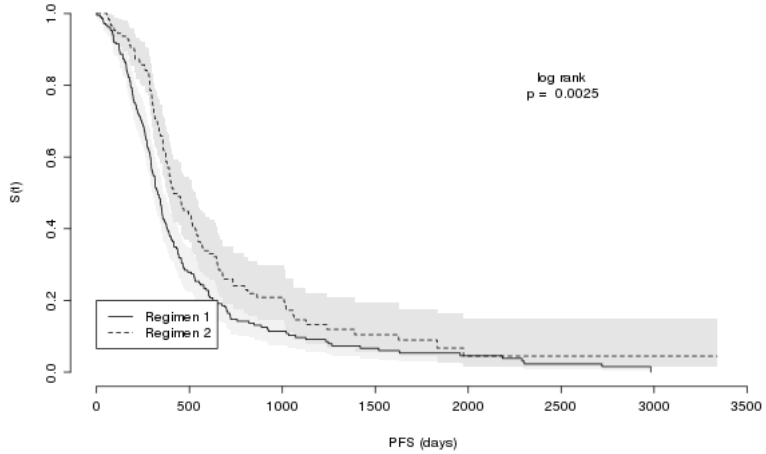


Figure 2.13: A survival function for the progression-free survival (PFS) for patients under different treatment regimen.

2.1.4.3 *Semi-parametric models*

Cox proportional hazards regression Cox proportional hazards regression models is one of the most popular mathematical models used for survival analysis; it was first described by D. R. Cox, its formula specifies the hazard function ($h(t)$) as follows [Cox, 1975, Collett, 2004b, Kleinbaum and Klein, 2005]:

$$h(t|x_i^{(j)}) = h_0(t)\exp(x_i^{(j)}\beta_i) \tag{2.1.4.18}$$

In analogy with logistic regression, where the logit is assumed to be represented by linear related predictors (see section 2.1.2.7 on page 53). A Cox proportional hazards survival model[35] assumes that $\ln(h(t))$, are linearly related to a set of predictors [Harrell. et al., 1996].

The examination of the influence of k co-variates on the survival time can be specified by a linear-like model of the log hazard.

$$\ln\left(h(t)\right) = \alpha(t) + \beta_1 x_1^{(j)} + \beta_2 x_2^{(j)} + \ldots + \beta_k x_k^{(j)} \tag{2.1.4.19}$$

$$h(t) = \exp\left(\alpha(t) + \beta_1 x_1^{(j)} + \beta_2 x_2^{(j)} + \ldots + \beta_k x_k^{(j)}\right) \tag{2.1.4.20}$$

$$h(t) = h_0(t)\exp\left(\sum_{i=1}^{k} x_i^{(j)}, \beta_j\right) \tag{2.1.4.21}$$

$$h(t) = h_0(t)\exp(f(X)) \tag{2.1.4.22}$$

The Cox proportional hazard function is a collection of two different parts [Collett, 2004b, Kleinbaum and Klein, 2005]: (1) *baseline hazard function* ($h_0(t)$) and (2) *relative risk function* [Harrell, 2012b]. An important characteristic of this formula is called the *proportional hazards assumption* [Kleinbaum and Klein, 2005]:

> The baseline hazard function is a function of time and is not related to one of the features, whereas the relative risk is independent of time and is related to the features.

[35] **R scripting**:
The following R libraries support the calculation of Cox proportional hazard survival models:

- survival: function coxph
- rms [Harrell, 2012a]: function cph

This proportional hazards assumption is very important for applying this regression models in the appropriate manner. What do we do with time-dependent features? Do they break the proportional hazards assumption? We will describe how to handle time-dependent features in the following sections. First, let's assume our features are time-independent. A Cox proportional hazards model with time-dependent features is often called an *extended Cox proportional hazards model*.

The baseline hazard rate ($h_0(t)$) is left unspecified, and therefore makes the Cox model *semi-parametric* [Kleinbaum and Klein, 2005, Harrell, 2012b]. According to the proportional hazards regression model, the baseline hazard function $h_0(t)$ is defined as the hazard function for an individual with all co-variates equal to zero[36]. This parameter can be nuisance, but in case we want to perform a prediction based on a Cox model, it can be estimated by the method of Breslow [Breslow, 1974].

[36] $h(t|x) = h_0(t)\exp(0)$

The relative risk function is the relationship between the features and the hazard function [Kalbfleisch and Prentice, 2002]. This relationship is in the Cox model is represented by the *partial likelihood* $\exp(x_i, \beta)$; Cox [Cox, 1975] estimates β by maximizing the partial likelihood (MLE). The name partial likelihood is derived from the independent non-informative censoring, i.e. the censored times are not taken into consideration.

$$L(\beta|x_i) = \prod_{i}^{n} \exp(x_i, \beta) \tag{2.1.4.23}$$

$$L(\beta) = \prod_{i=1}^{r} \frac{\exp(\beta' x_i)}{\sum_{k \in R_{(t_i)}} \exp(\beta' x_k)} \tag{2.1.4.24}$$

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(\beta' x_i)}{\sum_{k \in R_{(t_i)}} \exp(\beta' x_k)} \right)^{\delta_i} \tag{2.1.4.25}$$

$$\log(L(\beta)) = \sum_{i=1}^{n} \delta_i \left(\beta' x_i - \log \sum_{k \in R_{(t_i)}} \exp(\beta' x_k) \right) \tag{2.1.4.26}$$

- n individuals
- r distinct death times, and $n - r$ right censored survival times. The r ordered death times: $t_1 < t_2 < t_3, \ldots < t_r$. The likelihood function depends on the rank order of death times, the risk at each death time is determined.
- $R_{(t_i)}$ is the group of individuals are alive and uncensored at time just prior to t_j. This group is also called the *risk set*.
- x_i is the vector of co-variates for the individual that dies at the ith ordered death time t_i.

When the Cox proportional hazards regression parameters (β's) are estimated. The strength of different effects can be analysed

by the so-called *hazard ratio (HR)*. Two observations have corresponding linear estimates can be compared [Kleinbaum and Klein, 2005, Harrell, 2012b]:

$$o^{(j)} = \beta_1 x_1^{(j)} + \beta_2 x_2^{(j)} + \ldots + \beta_k x_k^{(j)} = \sum_{i=1}^{k} \beta_i X_i^{(j)} \tag{2.1.4.27}$$

$$o^{(j+1)} = \beta_1 x_1^{(j+1)} + \beta_2 x_2^{(j+1)} + \ldots + \beta_k x_k^{(j+1)} = \sum_{i=1}^{k} \beta_i X_i^{(j+1)} \tag{2.1.4.28}$$

The ratio of the two hazards is written as:

$$HR = \frac{h^{(j)}}{h^{(j+1)}} = \frac{h_0(t)e^{o^{(j)}}}{h_0(t)e^{o^{(j+1)}}} = \frac{e^{o^{(j)}}}{e^{o^{(j+1)}}} \tag{2.1.4.29}$$

$$= \exp\left(\sum_{i=1}^{k} \beta_i \left(X_i^{(j)} - X_i^{(j+1)}\right)\right) \tag{2.1.4.30}$$

The Cox proportional hazards regression model can plot the survival function based on the explanatory variables used as predictors. These survival curves are often called *adjusted survival curves*. The survival function can be written as:

$$S(t) = [S_0(t)]^{\exp(\Sigma_{i=1}^{k}\left(\beta_i^{(j)} x_i^{(j)}\right)} \tag{2.1.4.31}$$

An alternative way to formulate the proportional hazards assumption is to consider the hazard ratio. The hazard ratio of two observations is proportionality constant (α). It is crucial for applying the Cox proportional hazards regression to test this assumption[37].

$$\alpha = \exp\left(\sum_{i=1}^{k} \beta_i \left(X_i^{(j)} - X_i^{(j+1)}\right)\right) \tag{2.1.4.32}$$

[37] **R scripting**:
The following R libraries support the test of the proportional hazards assumption for the Cox model:
- survival: function cox.zph

Grambsch and Therneau [Grambsch and Terneau, 1994] have proposed a goodness-of-fit testing approach to test the proportional hazards assumption. This test is nowadays the standard test in various software packages. In following figure (see figure 2.14 on page 78), a result of the test is illustrated. The fitted line should be a constant line, otherwise the proportional hazards assumption is violated.

Interactions Interactions are product terms that will be added as a feature into a Cox proportional hazards regression model. In this type of regression models, the product terms are very often a function of time ($X_i^{(j)} z(t)$). The function $z(t)$ can have different forms (see table 2.17). Time-dependent features need to be used with caution because they easily violate the proportional hazards assumption.

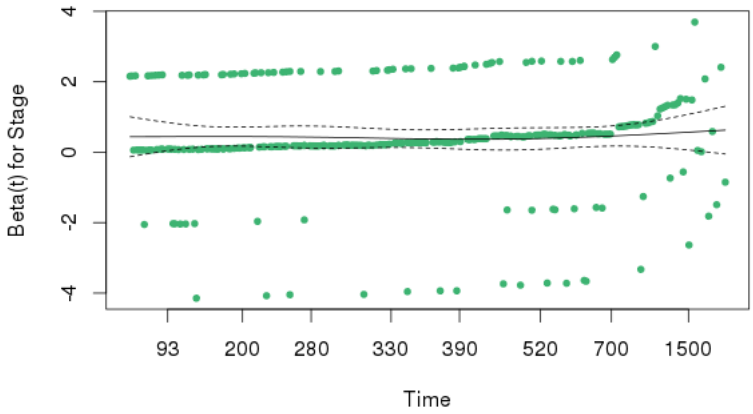


Figure 2.14: Analysis of the proportional hazards assumption with cox.zph function in R. The flatness of the fitted line illustrates that the Stage parameter does now violates the proportional hazards assumption over the survival time (Time).

When a feature violates the proportional hazards assumption, one possible solution is to use *stratification*. If this feature is continuous, it needs to be discretized. A stratified Cox proportional hazards regression model can be constructed with or without interactions. Without interactions, the feature is identified with strata in R, and can be formulated as:

$$\forall strat = 1, \ldots, s : h_{strat}(t) = h_{0strat}(t) \exp\left(\sum_{i=1}^{k} x_i^{(j)}, \beta_j\right) \quad (2.1.4.33)$$

A stratified Cox proportional hazards regression can also be constructed with interactions. In practice, it means that the model is fitted for different subsets of the sample. It can be formulated as:

$$\forall strat = 1, \ldots, s : h_{strat}(t) = h_{0strat}(t) \exp\left(\sum_{i=1}^{k} x_i^{(j)}, \beta_{j_{strat}}\right) \quad (2.1.4.34)$$

Table 2.17: Functions of time used as interactions in the Cox proportional hazards regression.

Function
$z(t) = t$
$z(t) = \log(t)$
$z(t) = \begin{cases} 0 & \text{if } t \geqslant t_0 \\ 1 & \text{if } t > t_0 \end{cases}$

2.1.4.4 *Feature selection*

Feature selection for a Cox proportional hazards regression model can be very similar as for a linear model, e.g. the Aikake information criterion, Bayesian information criterion, etc (see section 2.1.2.3 on page 51). Furthermore, there exist specific strategies to perform feature selection for Cox proportional hazards regression.

During this PhD, *supervised principal component analysis (SPCA)* [Bair and Tibshirani, 2004] and *least absolute shrinkage and selection operator (LASSO)* [Tibshirani, 1997] were used to perform feature selection for survival analysis.

Supervised principal component analysis (SPCA) Supervised principle component analysis applies *principal component analysis (PCA)*.

Principal Component Analysis (PCA) is a relatively old statistical method to identify patterns in data [Kendall, 2004]. This method is an eigen vector technique that denotes the variance in a data set [Hand et al., 2001].

A *principal component* is an *uncorrelated linear function* that represents the original data set [Kendall, 2004]. A data set with p variables and n samples $(X_1, X_2, X_3, \ldots, X_p)$, is centered to the mean of each variable [Holland, 2008]. This is required to insure that the principal components are centered in the data cloud without violating: spatial relationships in the data and variances along the variables [Holland, 2008].

The first principal component (PC_1) is a linear combination with stationary values, and expresses most of the variance in the data set [Smith, 2002]. In matrix notation PC_1 can be written [Holland, 2008]:

$$PC_1 = a_1^T X \tag{2.1.4.35}$$

Increasing the weights values (a_{ij}) can increase the variance of the first principal component, therefore the following constraint is introduced:

$$a_{11}^2 + a_{12}^2 + a_{13}^2 + \ldots + a_{1p}^2 = 1 \tag{2.1.4.36}$$

The second principal component (PC_2) is uncorrelated, or geometrically perpendicular, to the first principal component. The second principal component also represent the second highest variance a principal component can represent [Kendall, 2004] [Holland, 2008]. This continues to the pth principal component.

The most difficult question when PCA is applied: "Did a principal component contribute any meaningful information?". This question must be answered for each application of PCA. Supervised principal component analysis clusters the most important features by scoring the correlation to the output variable and constructs principal components for groups of co-expression.

$$y^{(j)} = \theta_0 + \theta_1 U + \epsilon \tag{2.1.4.37}$$

$$\forall i \in \mathcal{P} : X_i^{(j)} = \alpha_i + \beta_i U + \epsilon_i \tag{2.1.4.38}$$

Each feature (X_i) that is dependent on an underlying latent variable (U), will be used to estimate U and fit a model to predict the output variable $y^{(j)}$ [Bair et al., 2004]. As a result, this methodology can be used for features selection[38].

[38] **R package**:
The following R package support supervised principal component analysis:

- superpc [Bair and Tibshirani, 2004]: listfeatures outputs the selected features.

Least absolute shrinkage and selection operator (LASSO) The regression coefficients ($\hat{\beta}$) are estimated in a Cox proportional hazards regression model based on partial likelihood (see equation 2.1.4.26 on page 76). Tibshirani [Tibshirani, 1997] proposes an extra condition for computation of the regression coefficients [Tibshirani, 1995, Goeman, 2010]:

$$\log\left(l(\beta)\right) = \log\left(l(\beta)\right) - \lambda \sum_{i=1}^{k} |\beta_i| \tag{2.1.4.39}$$

$$\hat{\beta} = \arg\min\left[\log\left(L(\beta)\right)\right], \text{ subject to } \sum |\beta_i| \leqslant \lambda \tag{2.1.4.40}$$

LASSO computes the coefficients of a Cox proportional hazards model, so the sum of the absolute value of the coefficients are less or equal to a user-defined constant λ. Applying this extra condition for the optimization problem leads to less features in our resulting model[39].

Since a Cox proportional hazards regression model contains similarities with a linear model, a lot of the concepts used for under- and overfitting prevention in linear models also apply for the Cox proportional hazards regression (see section 2.1.2.4 on page 52).

In the following section, a performance measure specific for survival analysis will be defined.

[39] **R package**:
The following R package support LASSO for survival analysis:

- penalized [Goeman, 2010]: LASSO implementation based on gradient ascent and Newton-Raphson algorithm.
- glmpath [Park, 2007]: LASSO implementation based on path algorithm.

2.1.4.5 *Concordance index*

The *concordance index* [Harrell et al., 1982], also often called *c-index*, is a measurement of the discrimination capacity for a survival model [Gerds et al.,]. Censored data have specific constraints, which makes it more challenging to extract a predictive model. One of the main challenges is that many of the patients under investigation can live longer as the follow-up time of a study (right censoring, see 1 on page 73).

The c-index can be formulated as [Yan et al., 2004]:

$$c = \frac{\sum_{(p_i,p_j)\in\Omega} \hat{Y}\left(\hat{t}_i, \hat{t}_j\right)}{\Omega} \tag{2.1.4.41}$$

where

$$\hat{Y}(\hat{t}_i, \hat{t}_j) = \begin{cases} 1 & \text{if } \hat{t}_i > \hat{t}_j; \\ 0 & \text{otherwise} \end{cases}$$

a pair of patients (p_i, p_j)

all possible pairs of patients that can be classified Ω. In order to be a member of the collection of all possible pairs, the pair need to fulfill the following two conditions:

1. In case both patients of the pair (p_i, p_j), the event captured in the survival analysis reoccurred, and the recurrence time t_i of patient i is lower as the recurrence time t_j of patient j.
2. In case one patient experiences the event under investigation (p_i), and patient recurrence time t_i is shorter as the follow-up time of patient p_j.

prognostic scores from a survival model $\hat{t}_i, \hat{t}_j$

The c-index represents the probability that a patient with a higher prognostic survival score $(\hat{t}_i)$ will have a lower time to event as a

patient with a lower prognostic survival score ($\hat{f}_j$) [Harrell. et al., 1996] [Gerds et al.,].

The c-index needs to be interpreted in a specific manner. Its values reflect a specific meaning [Harrell. et al., 1996]: (a) if $c = 0.5$: no predictive discrimination of the patients with different outcomes, (b) $c = 0$ or $c = 1.0$ perfect discrimination of the patients with different outcomes.

As explained in the previous paragraphs, the c-index is a probability. Many clinicians are more used to a rank correlation coefficient ($[-1, +1]$). Therefore a Somers' D rank index was defined as:

$$D = 2(c - 0.5) \tag{2.1.4.42}$$

An alternative for Cox proportional hazards regression models will be introduced in the next section.

2.1.4.6 *Partial Cox regression (PCR)*

A partial Cox regression model is based on *Partial Least Squares (PLS)* [Garthwaite, 1994]. PLS is developed in 1960's by Herman Wold, and was first applied in econometrics. Nowadays it is applied in variety of fields: chemistry, monitoring and controlling industrial processes [Tobias, 2002]. PLS is an alternative for multiple linear regression (MLR) (see section 2.1.2 on page 47), and it can overcome the problem on overfitting in case the amount of observations is not bigger as the amount of explanatory variables. The effect of the explanatory variables will not be provided as explicitly as in the case of MLR, since PLS constructs underlying latent variables. Therefore, PLS is sometimes also called: *projection to latent structure* [Tobias, 2002].

The Partial Cox regression model can be written as [Li and Gui, 2004]:

$$h(t) = \exp\left(\alpha(t) + \beta_1 T_1 + \beta_2 T_2 + \ldots + \beta_k T_k\right) \tag{2.1.4.43}$$

$$h(t) = h_0(t) \exp(T_i, \beta_i) \tag{2.1.4.44}$$

$$h(t) = h_0(t) \exp(f(X)) \tag{2.1.4.45}$$

We will only explain the differences with 2.1.4.3 on page 75:

component T_k

risk function $f(X)$

Each component (T_j) and the risk function ($f(X)$) are a linear combination of all the different explanatory variables ($X = \{X_1, X_2, \ldots, X_k\}$). Taken into account the PLS principles for modelling the relationship between a explanatory variable X and the hazard. We restrict the relationship to the variable X has an influence on the hazard, and the other variables have no influence on this relationship. The other variables have an influence on the hazard by the

other components [Li and Gui, 2004]. The different components are computed sequentially, as for the first component we can write:

$$V_{1j} = X_j - \bar{x}_j \tag{2.1.4.46}$$

$$\bar{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n} \tag{2.1.4.47}$$

sample size n

vector of sample values V_{1j} $v_{1j} = \{v_{11j}, v_{21j}, \ldots, v_{n1j}\} = x_j - \bar{x}_j$. This implies that the mean of the samples of V_{1j} is equal to zero.

fit each variable in the following Cox model $h(t) = h_0(t)\exp(V_{1j}, \beta_{1j})$

2.1.5 *Resampling methods*

Resampling methods are applied for testing computational models in various manners. Bootstrapping and cross validation are very often used to test new constructed computational models.

Bootstrapping and Monte Carlo sampling are resampling methods that reconstruct a data sample in order to draw conclusion about the statistics of a computation model.

Cross validation is a statistical benchmark procedure that compares different learning algorithms by dividing the data into two segments: one for training and a second for validating. Cross validation is applied for classification and regression problems in machine learning [Larson, 1931]. Cross validation methodologies can specify achieve two highly related goals: (1) analysis of *generalization* of the algorithm and (2) comparison of the performance of different algorithms and/or parametrized models.

Next, bootstrapping and Monte Carlo resampling methods will be introduced. This will be followed with an overview of different performance measure for regression and classification models.

2.1.5.1 *Bootstrapping*

The bootstrap method was first introduced by B. Efron [Efron, 1979]. Nowaday, it has been applied in various applications: approximating the standard error of a sample estimate, estimation of the bias correction, confidence interval approximation, etc.

Bootstrapping will generate a new data sample from the current data sample. Such a newly generated data sample is often called *bootstrap subsample* or *phantom sample* [Davison and Hinkley, 1997]. This bootstrap sample is of equal samples size as the original sample size and will be used to reevaluate a performance measure, the mean value of a model parameter, or the error on a model parameter, etc.

A bootstrap method determines the data distribution from the data, without depending on the *Central Limit Theorem (CLT)*[40] [Davison and Hinkley, 1997].

The number of times the data needs to be resampled should be empirically derived [Davison and Hinkley, 1997].

[40] Central Limit Theorem is the most fundamental theorem in statistics. It states that a sample distribution can be approximated with a normal distribution.

2.1.5.2 *Monte Carlo sampling*

Monte Carlo sampling is an alternative approach for model testing. A new random data sample is generated based on the original data. This random data sample can be sampled from an fitted distribution, or the original data sample can be shuffled.

The Monte Carlo sampling techniques used during this doctorate compare a statistic (i.e. performance measure) of a computational model with a randomly generated data sample. A p-value quantifies where the statistic of a computational model is situated compared to the computed statistic distribution.

The number of times the data needs to be resampled should be empirically derived [Davison and Hinkley, 1997].

2.1.5.3 *Cross validation procedures*

Re-substitution validation In the re-substitution validation procedure all the available data is used to learn the model, and this same data is applied for testing. This procedure suffers a lot from *overfitting*.

Hold-out validation This type of procedure has two independent data sets for training and testing. It avoids an overlap of training and testing data.

K-fold cross validation Here the procedure starts to partition the data into k (nearly) equally sized segments, also called *folds*. The k-fold cross validation procedure runs with k equal the number of samples in the data. One observation from the original sample is used as the validation sample. The remaining samples are used as the training data.

Leave-one-out cross-validation (LOO-CV) This cross validation procedure is a special case of k-fold cross validation, where k is equal to one.

Repeated K-fold cross validation To obtain more reliable performance estimation or comparison, a large number of estimates are required. Therefore the k-fold cross validation can be executed multiple times.

2.1.5.4 *Performance measures for classification*

The formulation of different performance measures for classification will be provided in the following sections.

Confusion matrix A classification scheme can be analyzed by construction a *confusion matrix*. Such a matrix collects the counts of true positives (*TP*), true negatives (*TN*), false positives (*FP*), and false negatives (*FN*).

From the confusion matrix several performance measures are derived: *accuracy*, *sensitivity*, *specificity*, *positive predictive value*, and *negative predictive value*.

Table 2.18: A confusion matrix for the analysis of a classification model.

	$\hat{y}^{(j)} = 0$	$\hat{y}^{(j)} = 1$
$y^{(j)} = 0$	*TN*	*FN*
$y^{(j)} = 1$	*FP*	*TP*

Accuracy

$$acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1.5.1}$$

Sensitivity Sensitivity is also called *recall* or *true positive rate*.

$$sen = rec = \frac{TP}{TP + FN} \tag{2.1.5.2}$$

Specificity

$$spec = \frac{TN}{TN + FP} \tag{2.1.5.3}$$

False positive rate is equal to $1 - spec$.

Precision Precision is also called *positive predictive value*.

$$ppv = \frac{TP}{TP + FP} \tag{2.1.5.4}$$

Negative predictive value

$$npv = \frac{TN}{TN + FN} \tag{2.1.5.5}$$

Receiver operating characteristic (ROC) Receiver operating characteristic (ROC) is a detection measure that originally comes from signal detection theory. During the World War II radar images were used to distinct objects in images being an enemy target or not. The area under the ROC curve is a measure for the discriminatory capacity of a classification model [Metz, 1978, Fan et al., 2006][41].

Accuracy is a very naive measure for a classifier, i.e. if only 5% of the samples contain a state of a binary classifier, and the classifier would always result in the other state, the classifier would have a result of 95%. Therefore, specificity and sensitivity are more explanatory performance measures for a classifier. The ROC curve plots *sen* vs $1 - spec$.

[41] **R package**:
The following R package supports different performance measures:

- ROCR [Sing et al., 2005]: performance measures and visualization method for performance analysis of classifiers.

F-measure The F-measure is a measure that takes precision and recall into consideration:

$$F = 2 \times \frac{prec \times rec}{prec + rec} \tag{2.1.5.6}$$

SAR metric A more robust[42], combined performance measure is the *SAR metric* [Caruana and Niculescu-Mizil, 2004]. It combines squared error, accuracy, and area under the ROC curve.

[42] It is more robust because it combines more as one performance measure.

2.1.5.5 *Performance measures for regression*

The formulation of different performance measures for regression will be provided in the following sections. The estimated value in the formulations is written as: $\hat{y}$.

Error sum of squares (SSE)

$$SSE = \sum_{i=0}^{N} (E(y) - \hat{y})^2 \tag{2.1.5.7}$$

Total sum of squares (SST)

$$SST = \sum_{i=0}^{N} \left(\mu_y - \hat{y}\right)^2 \tag{2.1.5.8}$$

Estimated variance (MST)

$$MST = \frac{SST}{N-1} \tag{2.1.5.9}$$

Mean square error (MSE)

$$MSE = \frac{SSE}{N} \tag{2.1.5.10}$$

Coefficient of determination (R^2) The coefficient of determination is defined as the ratio between the sum of squares explained by the regression model (*SSE*) and the total sum of squares around the mean (*SST*).

$$R^2 = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST} \tag{2.1.5.11}$$

The coefficient of determination is also often explained as the ratio of the amount of variance captured by the regression model and the total amount of variance of the output variable. We can write:

$$\sigma_y = \sigma_{\hat{y}} + \sigma_{\hat{y}'} \tag{2.1.5.12}$$

$$\sum_{i=1}^{N} \left(y - \mu_y\right)^2 = \sum_{i=1}^{N} \left(\hat{y} - \mu_y\right)^2 + \sum_{i=1}^{N} (y - \hat{y})^2 \tag{2.1.5.13}$$

$$R^2 = \frac{\sum_{i=1}^{N} \left(\hat{y} - \mu_y\right)^2}{\sum_{i=1}^{N} \left(y - \mu_y\right)^2} \tag{2.1.5.14}$$

The coefficient of determination is also called *squared multiple correlation coefficient* [Harrell. et al., 1996], and can be alternatively be calculated as:

$$R^2 = 1 - \frac{(n-p)\,MSE}{(n-1)S_Y^2} \tag{2.1.5.15}$$

number of samples n

number of parameters p

sample variance of the response variable S_Y^2

Adjusted coefficient of determination is used to compensate for the fact that the sample is not the complete population sample. If

the data sample is the complete population sample, then there is no need for the adjusted coefficient of determination [Faraway, 2002c].

$$R^2_{adjusted} = 1 - \frac{MSE}{MST} \tag{2.1.5.16}$$

2.2 *Biological methodologies*

Recent developments in new technology for biological experiments lead to the collection of high-throughput, large-scale, and high quality data. On a transcriptome and genome-level microarrays [Schena et al., 1995], massively parallel sequencing (also called next-generation sequencing) [Schuster, 2008, Reis-Filho, 2009], etc., are nowadays well established technologies. The differentially expressed genes in microarray experiment can discover new targets for the composition of a biological footprint of a biological process, but can suffer from the sensitivity and specificity of the probes for a quantitative analysis. Next-generation sequencing will help to measure copy number aberrations, genetic aberrations, harbor mutations, etc., to potentially reclassify cancer types.

Cancer is a proteomic and genomic disease [Cesario and Marcus, 2011]. Genetic disorders active during tumour development can result in a phophorylation in the protein network. By measuring these phenotypes we are able to better diagnose and understand a cancer, have a better selection of biomakers for new drug development, and eventually result in personalized drug development.

However the success of comparative genomic hybridization (DNA) and transcription profiling (RNA) technologies result in phenotype discoveries. In order to answer the functional aftermath of these phenotypes, there is still a need for reliable proteomics characterization: translational regulation, post-translational modifications, etc. Therefore we need high-throughput proteomics technologies to study the structure, function, and activation of proteins. In the next sections, two proteomics technologies will be discussed: (1) *reverse phase protein array (RPPA)* and (2) *protein expression in tissue microarray (TMA)*.

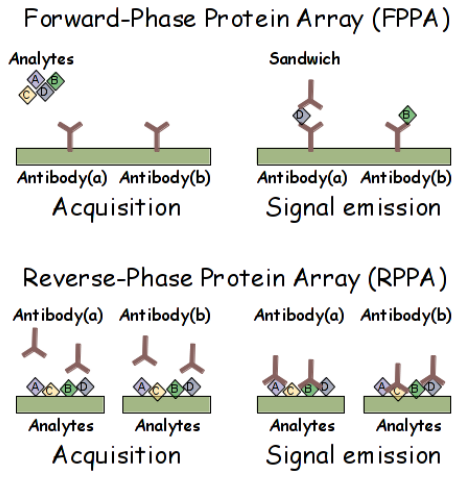


Figure 2.15: Forward phase protein array and reverse phase protein array have a different configuration of analytes and antibodies.

2.2.1 *Reverse phase protein array (RPPA)*

Reverse phase protein array (RPPA) are *protein microarray* [Lio, 2003]. A traditional microarray allows us to collect time series, with multiple measures for each time point, of an gene expression. Protein arrays are a grid of immunoblotting, i.e. western blotting, experiments. Experimental results show high correlation between RPPA and western blotting [Tibes et al., 2005].

There are two main types protein arrays: (1) *forward phase arrays* and (2) *reverse phase arrays* [Janzi et al., 2005, Spurrier et al., 2008]. The forward phase arrays have a *bait molecule*, i.e. antibody or antigen, that is present on the substratum of a spot. Each spot has one specific antibody and the array is incubated with one test sam-

ple, e.g. lysates of the condition of interest. A specific protein will bind directly or via a secondary antibody (i.e. sandwich) to the bait molecule, and multiple analytes are measured at once.

Reverse phase protein arrays has the analytes on the substratum of the spot, representing a single test sample, e.g. a patient sample. Each array will be incubated with a single antibody, and multiple combinations can be measured in parallel.

For maximizing precision and reliability and minimizing experimental variability, it is important to have predefined laboratory protocols, e.g. application of positive and negative controls [Spurrier et al., 2008] and apply computational methods to allow quantitatively interarray comparisons [Tibes et al., 2005].

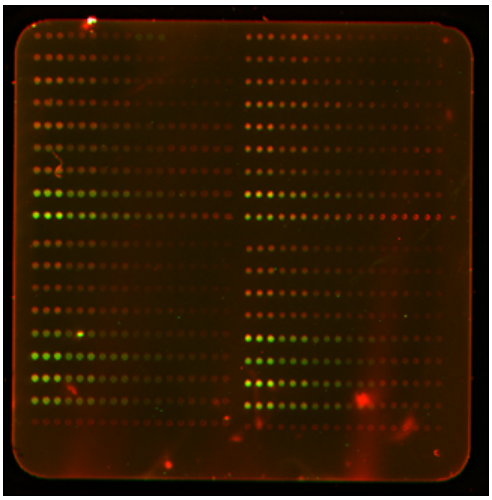


Figure 2.16: An example of an RPPA two-by-two grid plate. A set of 9 proteins are measure for a time-series with 17 intervals. Each dot on the figure represents the expression of an antibody of a corresponding target.

2.2.2 *Protein expression in tissue microarray (TMA)*

Tissue collection is performed by treating the biopsy with formalin (i.e. fixation) and paraffin, which results in a tissue block. With a microtome, a 5 μm slice of this residue is used to examine the tissue under the microscope. Kononen et al. [Kononen et al., 1998] introduced a *tissue microarray block* to assemble tissues with core needle biopsies of pre-existing tissue blocks [Camp et al., 2001]. These tissue microarrays are used to analyse tissue on the genome, transcriptome and proteome level by immunohistochemistry and immunofluorescence analysis [Kallioniemi et al., 2001].

This methodology allows to use archives of tissue blocks and collect "-omics" data in combination of clinicopathological data archives. In the following chapter 3, a study will be described combining these data resources for the prediction of overall- and progression-free survival.

Tumour tissues are a very complex mixture of malignant and benign tumour cells, stroma, extracellular material, etc., and can be an heterogeneous combination of different histological tumour types. Since a tissue on an array has a $2-5\mu$m diameter, it might not represent the true tumour tissue.

Automated quantitative analysis (AQUA®) uses a set of algorithms that allow quantitative protein expression of tissue microarrays [Camp et al., 2002]. This method is currently very popular in combination with immunofluorescence, and has shown to be the most efficient for the quantification of protein expression of tissue microarrays [Christopher B. Moeder and Rimm, 2009].

The classification of tumour and stroma in fluorescence tissue microarrays can improve the automatic quantitation of biomarker expression [Lahrmann et al., 2011].

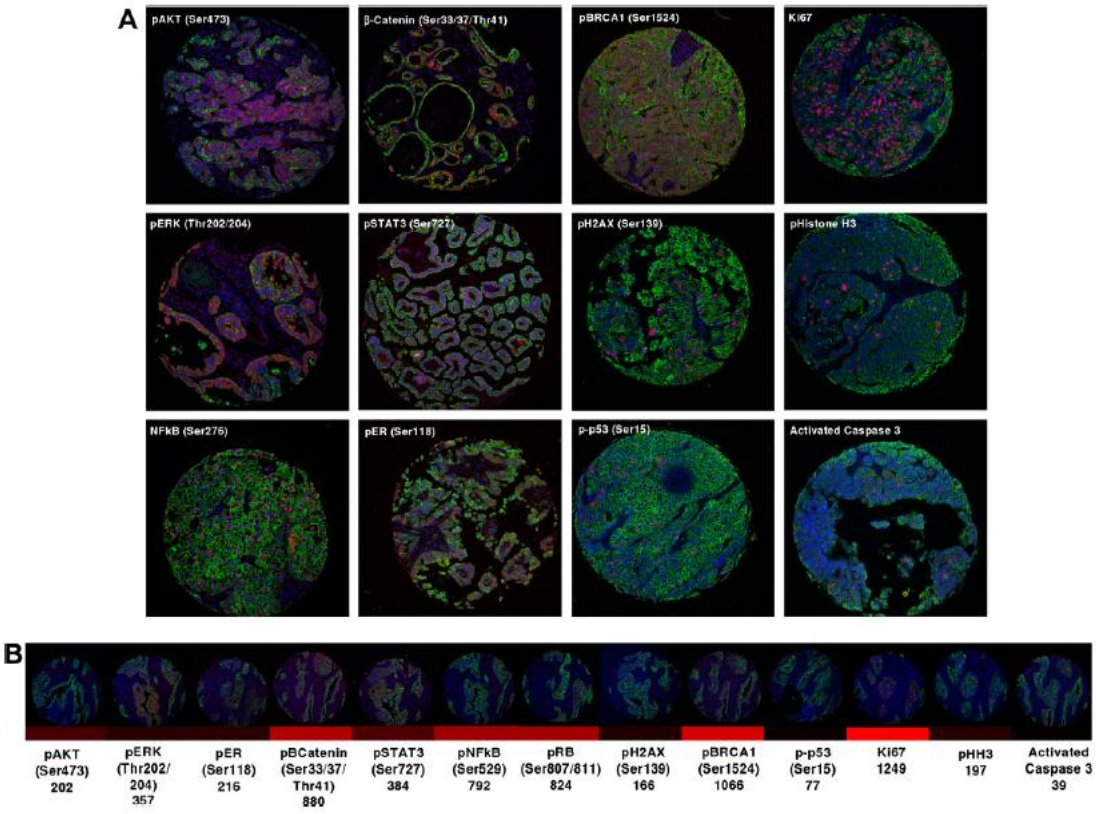


Figure 2.17: Immunofluorescence images of a tissue microarrays assay (Blue = DAPI nuclei; Green = cytokeratin tumour mask Red = antibody-conjugated flourophores) (From Fig. 1 of Faratian et al. [Faratian et al., 2011]).

3

Ovarian cancer

This chapter embargoed at the author's request.

4
Morphology during tumour invasion

> We must consider the distinctive characters and the general nature of plants from the point of view of their morphology, their behaviour under external conditions, their mode of generation, and the whole course of their life.
>
> *Theophrastus*

Studies in cancer biology often concentrate on the analysis of cancer cell and its genome; this type of research has identified numerous oncogenes and tumour-suppressor genes. New technologies have made a huge contribution in the collection and analysis of this data. One link that Systems Pathology aims to make is to map the heterogeneous and structurally complex nature of tissue and organ, called *tumour*, with this genome data. This is not only limited to genome data, but can be applied to the complete "-omics" scale.

Tumours are constantly on the move, they invade, communicate with their environment, and grow quickly. During this process they invade in mainly two modes: (1) *individual-* and *collective invasion*. For each of these invasion modes corresponding morphological measurements were performed with state-of-the-art image analysis.

A Bayesian network analysis was performed to learn the statistical dependencies between the modes of invasion and the morphological measures. These dependencies were further analysed to illustrated their discriminative properties. My contribution is this study is to provide analytical support for the determination of the discriminative capacity of morphological measurements for different types of tumour invasion.

My collaborators of the Division of Pathology at the University of Edinburgh collected the data of the tumour invasion. My main contribution is the analytical analysis; this was performed with a Bayesian network approach. The statistical dependencies found with the Bayesian network are independenty confirmed with a more traditional t-test. This research resulted in a publication [Katz et al., 2011].

4.1 Tumour invasion

One of the most complex phenomenon in cancer is *tumour invasion*. It is related to many programs of tumour development: *metastasis*, *epithelial-mesenchymal transition (EMT)*, *autophagy*, and *angiogenesis*. The environment of a tumour cell during invasion and migration, the interaction with this environment, the framework that forms the connective environment around a tumour cell, might play a crucial role in characterizing cancer and developing better drugs [Meuller and Fusenig, 2004].

The environment where cells are distributed is called *extracellular matrix (ECM)*, when a tumour exists its called *tumour matrix*. A tumour matrix facilitates a tumour to grow in at least three different tissue compartments [Cesario and Marcus, 2011]: (1) the original tumour compartment, (2) the mesenchyme of the primary site, also called *tumour invasion*, and (3) distant mesenchyme, also called *tumour metastatis*[1]. These different environments of tumour matrix can include specific cellular particles, e.g. blood-vessel cells, inflammatory cells, fibroblasts, and tissues related to wound healing.

[1] This is based in the definition of a tumour of Wallace H. Clark [Hong et al., 2010].

The inter-cellular communication among different types of cells is a fundamental program for tumour development. Tumours grow by excessive preparation of their environment, also called *tumour stroma*. One of the biggest problems with most cancer drugs is that cancer cells are precarious, and become resistant. Stroma cells are potentially better to treat, and a drug can normalize the stroma can block tumour development [Meuller and Fusenig, 2004, Elizabeth S et al., 2012].

One of the most fundamental programs in tumour invasion and migration is called *epithelial-mesenchymal transition (EMT)* (see 1.3 on page 23) [Friedl and Wolf, 2003, Hanahan and Weinberg, 2011]. This program, that incorporates a loss of epithelial biomarkers and a gain of mesenchymal biomarkers, illustrate that cancerous cells are able to reprogram their functionality [Yilmaz et al., 2007]. Katz et al. [Katz et al., 2010] have illustrated the overexpression of oncogene C35 leads to EMT-mediated invasion, which marks the transition between collective and individual invasion [Christiansen and Rajasekaran, 2006].

Obviously, tumours have other transition mechanisms for invasion and migration, i.e. *mesenchymal-amoeboid transition (MAT)* [Yilmaz et al., 2007]. And some of these programs are still not fully understood [Ilina and Friedl, 2009].

What is the difference between individual and collective invasion? In this study, different types of invasion are analysed with different morphological measures. Following the invasion model of Yilmaz et al. [Yilmaz et al., 2007], four different types of invasion are defined: type I: *single cell*, type II: *small group (2-5 cells/tumour)*, type III: *cohort (5-10 cells/tumour)*, and type IV: *coordinated (>10 cells/tumour)*. Type II is defined in vivo [Auguste et al., 2007]. Type I and type II are individual invasion, and type III and type IV are

collective invasion.

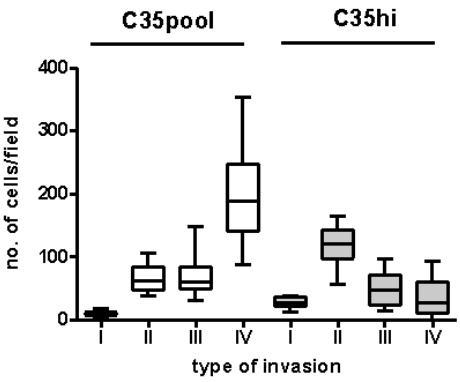


Figure 4.1: Boxplot for the different invasion types per cell line (C35pool and C35hi) [Katz et al., 2011].

Figure 4.2 pictures the result of the in vitro invasion model. This model is constructed with H16N-2 breast cells in a collagen lattice. The H16N-2 breast cells with C35 oncogene expression supports mammary epithelial cell invasion [Evans et al., 2006, Katz et al., 2010]. A cell line with variable intermediate levels of C35 expression results in mainly collective invasion (C35pool). Cells is high levels of C35 expression result in mainly individual invasion (C35hi) (see Fig. 4.1).

On top of each invasion assay (see figure 4.2), there is a group of cells with a very high length/width ratio. This groups is called the *origin-* or *seed group*. The origin group and every other each group of cells that has more as 22 % cell contact with this origin group was excluded from the analysis.

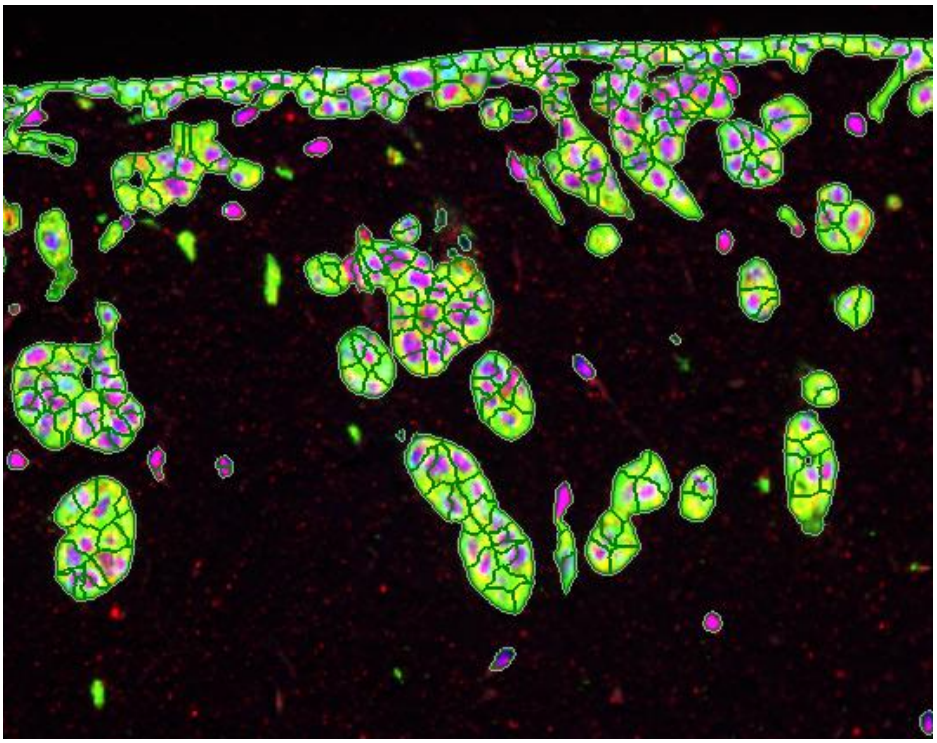


Figure 4.2: A fluorescent-stained image from invasion assay. Pan-cytokeratin rabbit polychonal antibody is used to select epithelial cells, and visualization is performed by anti-rabbit-Cy3. DAPI counterstain was used to identify nuclei.

4.2 *Automated image analysis*

The image analysis was performed with Definiens Cellenger ®. This software applies *cognition network technology (CNT)* for information extraction of invasion assays. The extraction uses knowledge-based and context-dependent processing which imitates human cognition [Athelogou et al., 2007]. This processing is defined in a functional programming language called *cognition network language (CNL)*.

The CNT-CNL facilitates a high-level semantic network analysis based on the two main steps: (1) *segmentation* and (2) *classification*.

Segmentation detects the *objects of interest* and classification depends on the objects of interest. Both steps are repeated which leads to an iterative process that performs the imaging processing [Baatz et al., 2009].

During the segmentation step, a hierarchical network is processed to detect the different objects of interest. In this study, the image pixels are first mapped into the *nucleus level*, where nuclei and cytoplasm are screened. Combining objects at the nuclei level leads to cell objects, and combining cell objects leads to the tumour level (see figure 4.3).

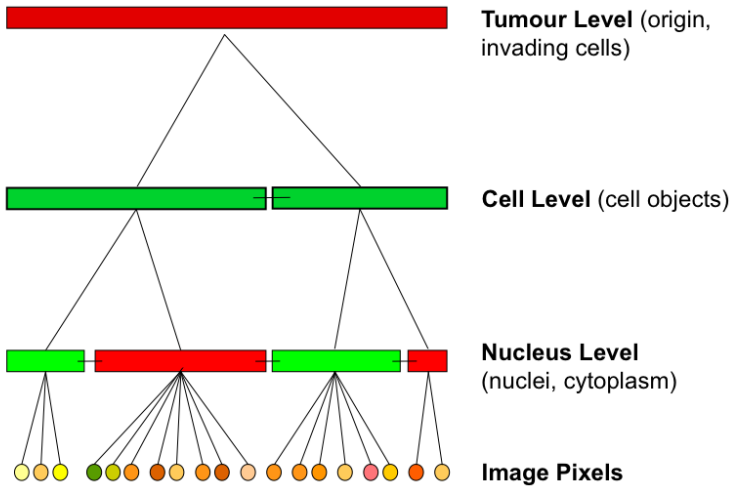


Figure 4.3: The cognition network technology (CNT) applied for the detection of tumours in the invasion assay.

4.2.1 *Morphological measures*

The CNT-based imaging analysis support the collection of morphological properties of the detected tumour objects. From the catalog of properties provide by Cell Cellenger ®, the following morphological measures were selected: *group area*, *roughness*, *roundness*, *length/width ratio*, *cell-cell contact*, and *border-to-origin*.

Group area, roundness, length/width ratio, roughness (called shape index in the Definiens Cellenger reference manual [AG., 2008]), and cell-cell contact do not need further explanation. The border-to-origin indicates the contact a group of cell contains with the origin group of cell at the top of each assay.

In the next sections, Bayesian networks analysis will be used to quantify and reason about the most important morphological measures during tumour invasion. It is expected that all these measures are important, but I would like to answer the question: "Which of these morphological measures are more important?".

4.3 *Bayesian network*

Bayesian networks are applied to learn the network of statistical dependencies between the morphological variables and tumour invasion types, and between morphological variables. Hereby, there are no statistical dependencies allowed between the mutual exclusive invasion types [Friedman, 2004].

The continuous morphological variables are discretized with quantile discretization (see 2.1.1.13 on page 44) with three discretization levels.

This discretized data will be used to construct a Bayesian network with a search and score approach. The BDeu scoring metric will be applied with a equivalent sample size equal to one. There is no dimension of time included into our data set, so a static Bayesian network is constructed. A greedy search algorithm is used with random restarts every 3000 iterations. The search will allow maximum three parent-child relationships following the constraints by the data amount (see table 2.11 on page 45) [Yu, 2005].

The statistical dependencies or links among the nodes in the Bayesian network are annotated with the influence score. This influence score, as described in 2.1.1.17 on page 46, illustrates the type of relationship. Red links, a positive influence score, indicate that high values in one node correspond to high values in the other node, and low with low. Blue links, a negative influence score, are the opposite, high values in one node correspond to low values in the other node, and low with high. Green links, when influence score is zero, indicate a non-monotonic relationship, e.g. U- or hump-shaped. The direction of the links in the Bayesian network are not included because they do not have any biological meaning.

The final Bayesian network is a consensus network based on the top 100 networks (see section 2.1.1.16 on page 46). This consensus network supports reasoning about the discriminative capacity morphological measures for the different invasion types.

The Bayesian network in figure 4.4 illustrates the statistical dependencies between the morphological measurements and the tumour invasion types. Cell-cell contact is linked to all invasion types, as expected the cell-cell contact has high discriminative power among the different types of invasion.

Group area is linked to the collective invasion types: cohort and coordinated, and roughness is linked to single cell and small group. Group area and roughness show to to be directly dependent to tumour invasion types.

The length/width ratio is not directly connected to any of the invasion types. This is an unexpected result. Since during tumour invasion, it is often thought that tumours elongate.

We examined the distributions of cell-cell contact, group area, roughness and length/width ratio for individual- and collective invasion per cell line (C35pool and C35hi).

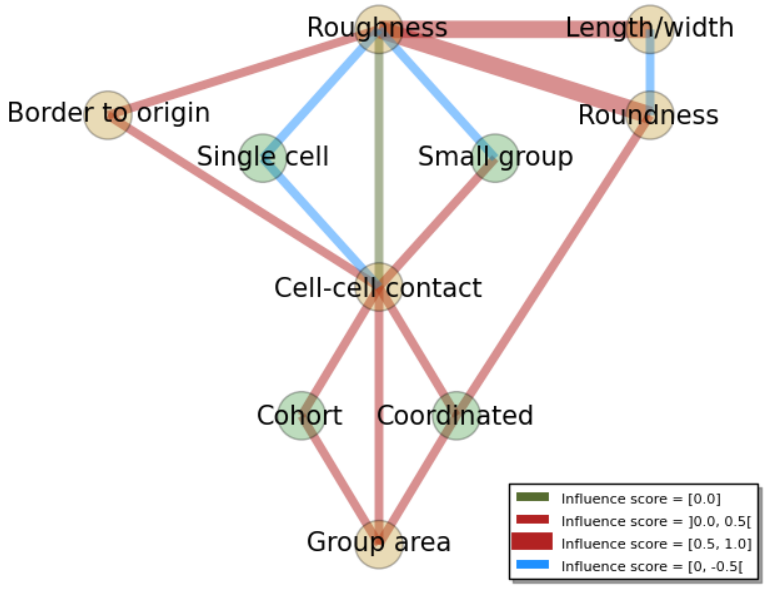


Figure 4.4: Bayesian network constructs a graph of statistical dependencies between morphological measures and tumour invasion types.

4.4 *Discriminative capacity of morphological measures*

A further statistical analysis is performed on the important morphological measures. A summary of the observed group objects are listed in the table 4.1.

Table 4.1: Summary of the number of objects analysis for each invasion type per cell line.

Number of group objects analysed		n		n
C35pool	individual	498	collective	290
C35hi	individual	732	collective	86

The observed values of cell-cell contact, group area, roughness and length/width ratio are listed in the following table 4.2. The mean and standard deviation as well as the Bonferroni-corrected p-values of a t-test are listed.

Table 4.2: Mean values and standard deviation (SD) for the morphological measurements during individual- and collective invasion.

		mean	SD		Mean	SD	p-value
Cell-cell-contact							
C35pool	individual	0.25	0.19	collective	0.65	0.08	$p < 0.001$
C35hi	individual	0.17	0.15	collective	0.53	0.11	$p < 0.001$
Group area							
C35pool	individual	355.81	255.78	collective	1956.52	1243.44	$p < 0.001$
C35hi	individual	236.40	176.51	collective	1085.98	575.03	$p < 0.001$
Roughness							
C35pool	individual	1.335	0.189	collective	1.543	0.272	$p < 0.001$
C35hi	individual	1.392	0.179	collective	1.712	0.201	$p < 0.001$
Length/width ratio							
C35pool	individual	1.68	0.64	collective	1.79	0.67	$p = 0.15$
C35hi	individual	1.73	0.57	collective	1.72	0.58	$p = 1.0$

4.4.1 *Cell-cell contact*

Cell-cell contact is significant higher in collective invasion compared to individual invasion ($p < 0.001$). In the Bayesian network, cell-cell contact is the only morphological variable connected to all invasion types.

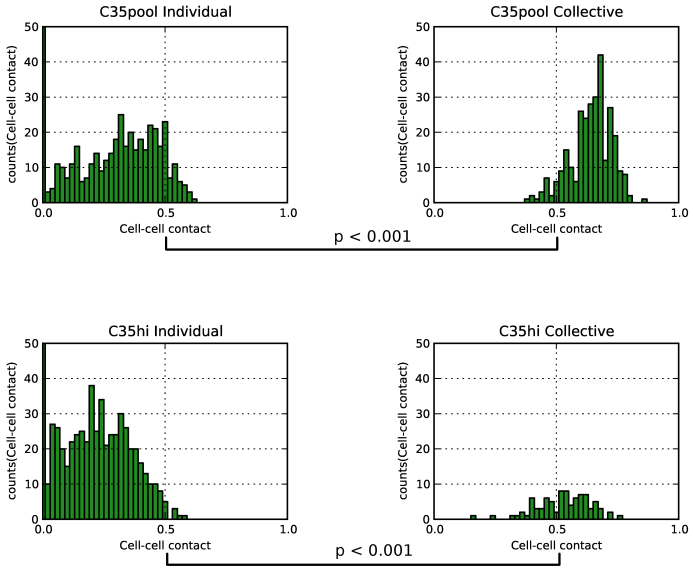


Figure 4.5: Histogram cell-cell contact for the comparison between individual- and collective invasion.

4.4.2 *Group area*

Group area is significant higher in collective invasion compared to individual invasion ($p < 0.001$). The area of cells during individual invasion does not compensate for the area of more cells in collective invasion, more cells still remain to have a bigger area for group objects.

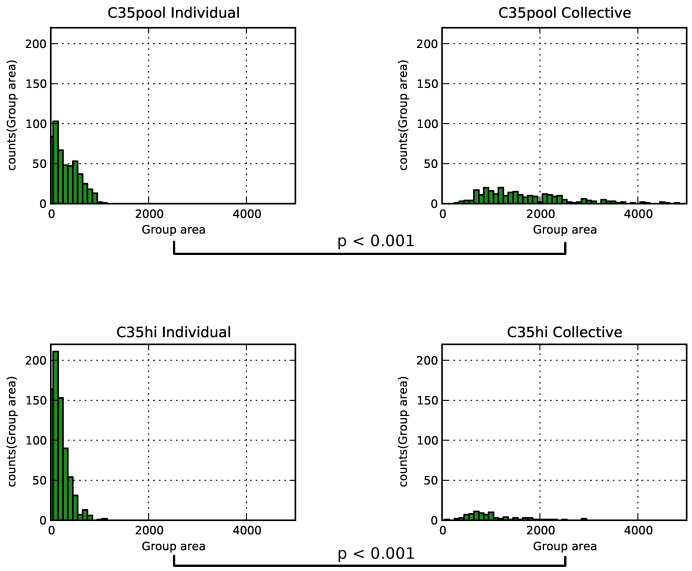


Figure 4.6: Histogram group area for the comparison between individual- and collective invasion.

4.4.3 *Surface roughness*

Individual invasion (single cell and small group objects) occurs in general with a smoother surface as collective invasion (cohort and coordinated). The roughness of collective invasion can be of different sources. One possible source of increased roughness is the formation of *invadopodia* [Weaver, 2006]. Invadopodia is a biological process of the interaction between membrane protrusions (i.e., extensions of the cell membrane) and proteolysis of the extracellular matrix (i.e., breakdown of proteins into polypeptide or amino acids) [Schoumacher et al., 2010]. A small experiment illustrated collective invasive structures expressed vimentin and MT1-MMP, both are documented to be active invadopodia [Attanasio et al., 2011].

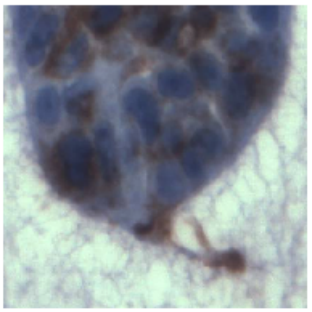


Figure 4.7: Invadopodia are a source of surface roughness during collective invasion. Example of vimentin staining of a collective invasive group [Katz et al., 2011].

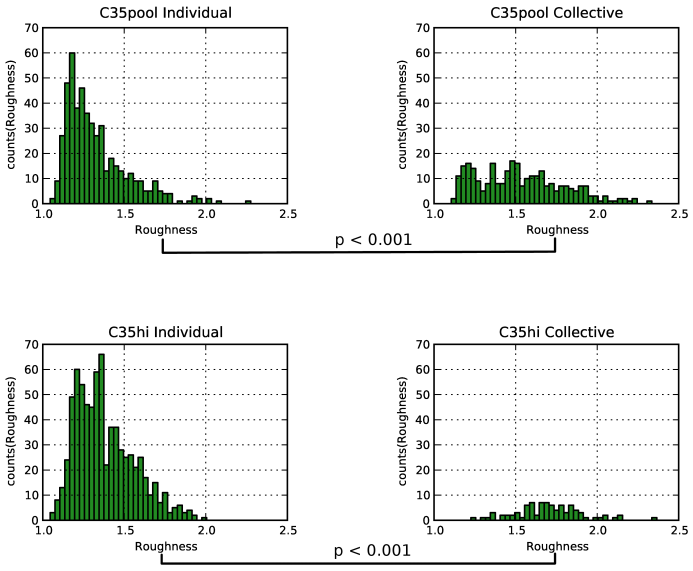


Figure 4.8: Histogram roughness for the comparison between individual- and collective invasion.

4.4.4 *Length/width ratio*

Length/width ratio illustrated to have no discriminative capacity between individual- and collective invasion. As a consequence, elongation of invasive groups is not only occurring in individual invasion; the same phenomenon is observed in the Drosophila ovary [Wang et al., 2010].

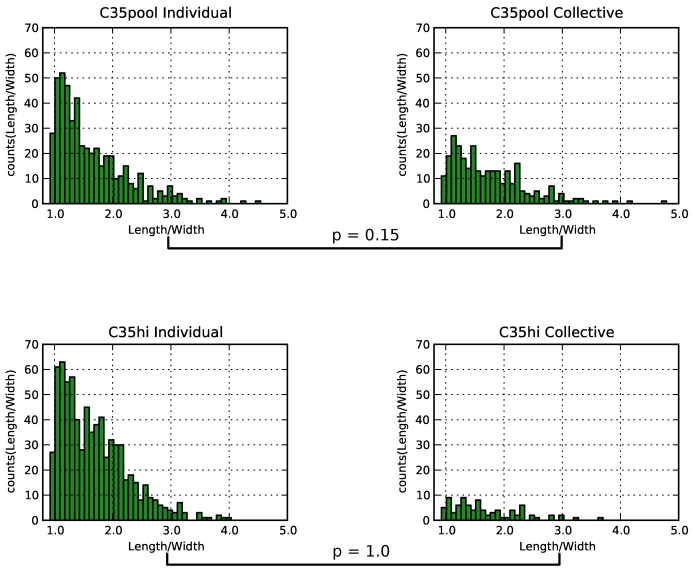


Figure 4.9: Histogram length/width ratio for the comparison between individual- and collective invasion.

4.5 *Discussion*

Cognition network technology (CNT) based imaging analysis was able to collect data of tumour invasion. This method was successfully used by the collaborators of the Division of Pathology at the University of Edinburgh. The analytical analysis that I performed allowed us to draw following conclusions.

Cell-Cell contact, group area, and surface roughness have the most discriminative capacity between individual- and collective invasion. In order to come to that conclusion, I performed a Bayesian network analysis. The Bayesian network allows us to integrate the tumour invasion types (i.e., single cell, small group (2-4 cells/tumour), cohort (5-10 cells/tumour), and coordinated (>10 cells/tumour)) as nodes into the network; the resulting network (see figure 4.4 on page 132) pictures the most prominent statistical dependencies among the nodes, the tumour invasion types and the morphological variables. An independent statistical t-test confirms the statistical dependencies found in the Bayesian network.

Cell-cell contact is the only morphological measure that is connected to all tumour invasion types in the Bayesian network.

An additional variable highlighted by Bayesian network analysis is that only collectively invading groups (>5 cells) are similarly

linked to the overall group area. Similarly, roughness is only connected to the individually invading groups (<5 cells).

The length/width ratio has less discriminative capacity between individual- and collective invasion as presumed. It suggest that the elongation of a tumour occurs an equal amount of degree in individual invasion as in collective invasion [Friedl and Wolf, 2003].

This research shows two distinct modes of invasion are found. Morphological differences between those invasion modes could now be exploited in both organotypic cell lines models and human cancer specimens to measure changes in tumour progression or drug response. Future work could research how organotypic models such as this used here could be directly translated to the models in a vivo setting [Timpson et al., 2011].

5

Conclusion

This chapter embargoed at the author's request.

6
Bibliography

[Lio, 2003] (2003). Protein microarrays: Meeting analytical challenges for clinical application. *Cancer Cell*, 3:317–325.

[Med, 2008] (2008). Online virtual medical center. http://www.virtualmedicalcenter.com/.

[EUD, 2012] (2012). Causes of death statistics.

[AG., 2008] AG., D. (2008). *Definiens Developer XD 1.1 – Reference Book*. Definiens.

[Akaike, 1974] Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automated control*, AC-16-6:716–723.

[Albert et al., 2002] Albert, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). *Molecular Biology of the Cell*. Garland Science, fourth edition.

[Athelogou et al., 2007] Athelogou, M., Schmidt, G., Schäpe, A., Baarz, M., and Binnig, G. (2007). Cognition network technology – a novel multimodal image analysis technique for automatic identification and quantification of biological image contents. *Imaging Cellular and Molecular Biological Functions*, 3:407–421.

[Attanasio et al., 2011] Attanasio, F., Caldieri, G., Giacchetti, G., van Horssen, R., Wieringa, B., and Buccione, R. (2011). Novel invadopodia components revealed by differential proteomic analysis. *European Journal of Cell Biology*, 90:115–127.

[Auguste et al., 2007] Auguste, P., Fallavollita, L., Wang, N., Burnier, J., Bikfalvi, A., and Brodt, P. (2007). The hostinflammatory response promotes liver metastasis by increasing tumor cell arrest and extravasation. *The American Journal of Pathology*, 170:1781–1792.

[Azim and Jr., 2008] Azim, H. and Jr., H. A. A. (2008). Targeting her-2/neu in breast cancer: As easy as this! *Oncology*, 74:150–157.

[Baatz et al., 2009] Baatz, M., Zimmermann, J., and Blackmore, C. G. (2009). Automated analysis and detailed quantification of

biomedical images using definiens cognition network technology®. *Combinatorial Chemistry and High Throughput Screening*, 12:908–916.

[Bair et al., 2004] Bair, E., Hastie, T., Debashis, P., and Tibshirani, R. (2004). Prediction by supervised principal supervised component analysis. Technical report, Standord University.

[Bair and Tibshirani, 2004] Bair, E. and Tibshirani, R. (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biology*, 2:511–522.

[Baldi and Brunak, 2001] Baldi, P. and Brunak, S. (2001). *Bioinformatics: the machine learning approach*. MIT Press, second edition.

[Barabási, 2003] Barabási, A.-L. (2003). *Linked*, pages 79–122. Plume, first edition.

[Barabási and Albert, 1999] Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286:509–512.

[Barabási and Oltai, 2004] Barabási, A.-L. and Oltai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nature: reviews Genetics*, 5.

[Barabási et al., 2011] Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12:56–68.

[Basak et al., 2007] Basak, D., Pal, S., and Patranabis, D. C. (2007). Support vector regression. *Neural Information Processing*, 11:203–224.

[Beck et al., 2011] Beck, A. H., Sangoi, A. R., Leung, S., Marinelli, R. J., Nielsen, T. O., van de Vijver, M. J., West, R. B., van de Rijn, M., and Koller, D. (2011). Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science Translational Medicine*, 3(108):108–113.

[Bishop, 2006a] Bishop, C. M. (2006a). *Pattern Recognition and Machine Learning*. Springer Science, first edition.

[Bishop, 2006b] Bishop, C. M. (2006b). *Pattern Recognition and Machine Learning*, pages 196–212. Springer Science, first edition.

[Bishop, 2006c] Bishop, C. M. (2006c). *Pattern Recognition and Machine Learning*, pages 1–66. Springer Science, first edition.

[Bishop, 2006d] Bishop, C. M. (2006d). *Pattern Recognition and Machine Learning*, pages 325–356. Springer Science, first edition.

[Bongard and Lipson, 2007] Bongard, J. and Lipson, H. (2007). Automated reverse engineering of nonlinear dynamical systems. *PNAS*, 104:9943–9948.

[Boser et al., 1992] Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Annual ACM Workshop on Computational Learning Theory*, pages 144–152.

[Bottcher and Dethlefsen., 2009] Bottcher, S. G. and Dethlefsen., C. (2009). *deal: Learning Bayesian Networks with Mixed Variables*. R package version 1.2-33.

[Brenner, 2010] Brenner, S. (2010). Sequences and consequences. *Philosophical transactions of the Royal Society Biological Sciences*, 365:207–212.

[Breslow, 1974] Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, 30:89–99.

[Burges, 1998] Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledges Discovery*, 2:121–167.

[Butcher et al., 2004] Butcher, E. C., Berg, E. N., and Kunkel, E. J. (2004). Systems biology in drug discovery. *Nature: Biotechnology*, 22:1253–1259.

[Butterworth et al., 2004] Butterworth, R., Simovici, D. A., Santos, G. S., and Ohno-Machado, L. (2004). A greedy algorithm for supervised discretization. *Elsevier Science*.

[Camp et al., 2001] Camp, R. L., Charette, L. A., and Rimm, D. L. (2001). Validation of tissue microarray technology in breast carcinoma. *Laboratory Investigation*, 80:1943–1949.

[Camp et al., 2002] Camp, R. L., Chung, G. G., and Rimm, D. L. (2002). Automated subcellular localization and quantification of protein expression in tissue microarrays. *Nature Medicine*, 8:1323–1327.

[Canty and Ripley, 2012] Canty, A. and Ripley, B. (2012). Package boot. *R package version 1.3-4*.

[Carmeliet, 2005] Carmeliet, P. (2005). Angiogenesis in life, disease and medicine. *Nature*, 438:932–936.

[Caruana and Niculescu-Mizil, 2004] Caruana, R. and Niculescu-Mizil, A. (2004). Data mining in metric space: An empirical analysis of supervised learning performance criteria. *ROCAI*, pages 9–18.

[Celis et al., 2009] Celis, J. E., Cabezón, T., Moreira, J. M., Gromov, P., Gromova, I., Timmermans-Wielenga, V., Iwase, T., Akiyama, F., Honma, N., and Rank, F. (2009). *Molecular Oncology*, 3:220–237.

[Cerny, 1985] Cerny, V. (1985). Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, 45:41–51.

[Cesario and Marcus, 2011] Cesario, A. and Marcus, F. (2011). *Cancer Systems Biology, Bioinformatics and Medicine*. Springer.

[Chan, 2001] Chan, J. K. C. (2001). The new world health organization classification of lymphomas: the past, the present and the future. *Hematological Oncology*, 19:129–150.

[Chang and Lin, 2011] Chang, C.-C. and Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2.

[Cheang et al., 2006] Cheang, M. C., Voduc, D., Bajdik, C., Leung, S., McKinney, S., Chia, S. K., Perou, C. M., and Nielsen, T. O. (2006). Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype. *Clinical Cancer Research*, 14:1368–1376.

[Chen et al., 2005] Chen, P.-H., Lin, C.-J., and Scholköpf, B. (2005). A tutorial on ν-support vector machines. *Applied Stochastic Models in Business and Industry*, 21:111–136.

[Chen and Lin, 2006] Chen, Y.-W. and Lin, C.-J. (2006). *Combining SVMs with Various Feature Selection Strategies*, pages 315–324. Springer.

[Cheng et al., 1997] Cheng, J., Bell, D. A., and Liu, W. (1997). Learning belief networks from data: An information theory based approach. *Proceedings of the sixth international conference on Information and knowledge management*, pages 325–331.

[Chickering, 1996] Chickering, D. M. (1996). *Learning Bayesian Networks is NP-Complete*, pages 121–130. Springer.

[Chickering, 2003] Chickering, D. M. (2003). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554.

[Christiansen and Rajasekaran, 2006] Christiansen, J. J. and Rajasekaran, A. K. (2006). Reassessing epithelial to mesenchymal transition as a prerequisite for carcinoma invasion and metastatis. *Cancer Research*, 66:8319–8326.

[Christopher B. Moeder and Rimm, 2009] Christopher B. Moeder, Jennifer M. Giltnane, S. P. M. and Rimm, D. L. (2009). Quantitative, fluorescence-based in-situ assessment of protein expression. *Methods in Molecular Biology*, 520:163–175.

[Clyde, 2006] Clyde, R. G. (2006). *Introduction - thesis*. PhD thesis, University of Abertay Dundee; 5-38.

[Clyde et al., 2006] Clyde, R. G., Bown, J. L., Zhelev, N., and Crawford, J. W. (2006). The role of modelling in identifying drug targets for diseases of the cell cycle. *The Royal Society*, 3:617–627.

[Collett, 2004a] Collett, D. (2004a). *Modelling survival data in medical research*, pages 15–55. Chapman & HALLCRC, second edition.

[Collett, 2004b] Collett, D. (2004b). *Modelling survival data in medical research*. Chapman & HALLCRC, second edition.

[Cooper and Herskovits, 1992] Cooper, G. F. and Herskovits, E. (1992). A bayesian method for the induction of probabilistic networks from data. *Journal of Machine Learning*, 9:309–347.

[Copas, 1983] Copas, B. (1983). Regression, prediction and shrinkage. *Journal of the Royal Statistical Society*, 45:311–354.

[Cowell, 2001] Cowell, R. G. (2001). Conditions under which conditional independence and scoring methods lead to identical selection of bayesian network models. *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 91–97.

[Cox, 1975] Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

[Crisp and Burges, 1999] Crisp, D. J. and Burges, C. J. (1999). A geomertic interpretation of ν-svm clssifiers. *Neural Information Processing Systems*.

[Csete and Doyle, 2002] Csete, M. E. and Doyle, J. C. (2002). Reverse engineering of biological complexity. *Science*, 295:1664–1669.

[Daly and Shen, 2009] Daly, R. and Shen, Q. (2009). Learning bayesian network equivalence classes with ant colony optimization. *Journal of Artificial Intelligence Research*, 35:391–447.

[Daly et al., 2009] Daly, R., Shen, Q., and Aitken, S. (2009). Learning bayesian networks: approaches and issues. *The Knowledge Engineering Review*.

[Dancey et al., 2010] Dancey, J. E., Dobbin, K. K., Groshen, S., Jessup, J. M., Hruszkewycz, A. H., Koehler, M., Parchment, R., Ratain, M. J., Shankar, L. K., Stadler, W. M., True, L. D., Gravell, A., and Grever, M. R. (2010). Guidelines for the development and incorporation of biomarker studies in early clinical trials of novel agents. *Clinical Cancer Research*, 16:1745–1755.

[Davidson et al., 2012] Davidson, B., Trope, C. G., and Reich, R. (2012). Epithelial-mesenchymal transition in ovarian carcinoma. *Frontiers in Oncology*, 2:1–13.

[Davison and Hinkley, 1997] Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*, pages 11–79. Cambridge Press.

[de Campos, 2006] de Campos, L. M. (2006). A scoring function for learning bayesian networks based on mutual information and conditional independence tests. *Journal of Machine Learning Research*, 6:2149–2187.

[Domingos, 1999] Domingos, P. (1999). The role of occam's razor in knowledge discovery. *Data Mining and Knowledge Discovery*, 3:409–425.

[Edge and Compton, 2010] Edge, S. B. and Compton, C. C. (2010). The american joint committee on cancer: the 7th edition of the ajcc cancer staging manual and the future of tnm. *Annals if Surgical Oncology*, 17:1471–1474.

[Efron, 1979] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7:1–26.

[Elizabeth S et al., 2012] Elizabeth S, N., Askautrud, H. A., Kees, T., Park, J.-H., Plaks, V., Ewald, A. J., Fein, M., Rasch, M. G., Qiu, Y.-X., Park, J., Sinha, P., Bissell, M. J., Frengen, E., Werb, Z., and Egeblad, M. (2012). Imaging tumor-stroma interactions during chemotherapy reveals contributions of the microenvironment to resistance. *Cancer Cell*, 21:488–503.

[Evans et al., 2006] Evans, E. E., Henn, A. D., Jonason, A., Paris, M. J., Schiffhauer, L. M., Borrello, M. A., Smith, E. S., Sahasrabudhe, D. M., and Zauderer, M. (2006). C35 (c17orf37) is a novel tumor biomarker abundantly expressed in breast cancer. *Molecular Cancer Therapeutics*, 5:2919–2930.

[Fan et al., 2006] Fan, J., Upadhye, S., and Worster, A. (2006). Understanding receiver operating characteristic (roc) curves. *Canadian Journal of Emergency Medicine*, 8:19–20.

[Faratian and Bartlett, 2008] Faratian, D. and Bartlett, J. (2008). Predictive markers in breast cancer – the future. *Histopathology*, 52:91–98.

[Faratian et al., 2009] Faratian, D., Clyde, R. G., Crawford, J. W., and Harrison, D. J. (2009). Systems pathology: taking molecular pathology into a new dimension. *Nature Reviews Clinical Oncology*, 6:455–464.

[Faratian et al., 2011] Faratian, D., Um, I., Wilson, D. S., Mullen, P., Langdon, S. P., and Harrison, D. J. (2011). Phosphoprotein pathway profiling of ovarian carcinoma for the identification of potential new targets for therapy. *European Journal of Cancer*, 47:1420–1431.

[Faraway, 2002a] Faraway, J. J. (2002a). *Practical Regression and ANOVA using R*, pages 36–40. R project, first edition.

[Faraway, 2002b] Faraway, J. J. (2002b). *Practical Regression and ANOVA using R*, pages 65–71. R project, first edition.

[Faraway, 2002c] Faraway, J. J. (2002c). *Practical Regression and ANOVA using R*, pages 129–130. R project, first edition.

[Fass, 2008] Fass, L. (2008). Imaging and cancer: A review. *Molecular Oncology*, 2:115–152.

[Fayyad and Irani, 1993] Fayyad, U. M. and Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. pages 1–6.

[Fidler, 2003] Fidler, I. J. (2003). The pathogenesis of cancer metastasis: the seed and soil hypothesis revisited. *Nature Reviews Cancer*, 3:453–458.

[Folkman, 1995] Folkman, J. (1995). Angiogenesis in cancer, vascular, rheumatoid and other disease. *Nature Reviews: Cancer*, 1:27–31.

[Friedl and Wolf, 2003] Friedl, P. and Wolf, K. (2003). Tumour-cell invasion and migration: Diversity and escape machanisms. *Nature Reviews: Cancer*, 3:362–374.

[Friedman, 2004] Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science*, 303:799–805.

[Fu, 1998] Fu, W. J. (1998). Penalized regression: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7:397–416.

[Garnett et al., 2012] Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., Greninger, P., Thompson, I. R., Luo, X., Soares, J., Liu, Q., Iorio, F., Surdez, D., Chen, L., Milano, R. J., Bignell, G. R., Tam, A. T., Davies, H., Stevenson, J. A., Barthorpe, S., Lutz, S. R., Kogera, F., Lawrence, K., McLaren-Douglas, A., Mitropoulos, X., Mironenko, T., Thi, H., Richardson, L., Zhou, W., Jewitt, F., Zhang, T., O/'Brien, P., Boisvert, J. L., Price, S., Hur, W., Yang, W., Deng, X., Butler, A., Choi, H. G., Chang, J. W., Baselga, J., Stamenkovic, I., Engelman, J. A., Sharma, S. V., Delattre, O., Saez-Rodriguez, J., Gray, N. S., Settleman, J., Futreal, P. A., Haber, D. A., Stratton, M. R., Ramaswamy, S., McDermott, U., and Benes, C. H. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483:570–578.

[Garthwaite, 1994] Garthwaite, P. H. (1994). An interpretation of partial least squares. *Journal of the American Statistical Association*, 20:208–215.

[Geiger et al., 1990] Geiger, D., Verma, T., and Pearl, J. (1990). Indentifying independence in bayesian networks. *Networks*, 20:507–534.

[Gelman et al., 2004a] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004a). *Bayesian Data Analysis*, pages 75–103. Chapman & HALLCRC, second edition.

[Gelman et al., 2004b] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004b). *Bayesian Data Analysis*. Chapman & HALLCRC, second edition.

[Gerds et al.,] Gerds, T. A., Kattan, M. W., Schumacher, M., and Yu, C. Estimating a time-dependent concordance index for survival

prediction models with covariate dependent censoring. *Statistics in Medicine*.

[Getoor and Taskar, 2007] Getoor, L. and Taskar, B. (2007). *Introduction to Statistical Relational Learning*, pages 175–200. The MIT Press, first edition.

[Goeman, 2010] Goeman, J. J. (2010). l_1 penalized estimation in the cox proportional hazards model. *Biometrical Journal*, 52:70–84.

[Goldman and Yang, 2008] Goldman, N. and Yang, Z. (2008). Introduction. statistical and computational challenges in molecular phylogenetics and evolution. *Philosophical transactions of the Royal Society Biological Sciences*, 363:3889–3892.

[Gonzalez et al., 2008] Gonzalez, L. O., Corte, M. D., Vazquez, J., Junquera, S., Sanchez, R., Alvarez, A. C., Rodriguez, J. C., Lamelas, M. L., and Vizoso, F. J. (2008). Androgen receptor expresion in breast cancer: Relationship with clinicopathological characteristics of the tumors, prognosis, and expression of metalloproteases and their inhibitors. *BMC Cancer*, 8:n/a–n/a.

[Grafen and Hails, 2006] Grafen, A. and Hails, R. (2006). *Modern Statistics for the Life Sciences*. Oxford University Press, first edition.

[Grambsch and Terneau, 1994] Grambsch, P. M. and Terneau, T. M. (1994). Proportional hazards test and diagnosis based on weighted residuals. *Biometrika*, 81:515–526.

[Gray and Druker, 2012] Gray, J. and Druker, B. (2012). The breast cancer landscape. *Nature*, 486:328–329.

[Greene and Sobin, 2009] Greene, F. L. and Sobin, L. H. (2009). A worldwide approach to the tnm staging system: Collaborative efforts of the ajcc and uicc. *Journal of Surgical Oncology*, 99:269–272.

[Grzegorczyk, 2006] Grzegorczyk, M. (2006). *Comparative Evaluation of Different Graphical Models for the Analysis of Gene Expression Data*. PhD thesis, University Dortmund; 25-74, Dortmund.

[Grzegorczyk and Husmeier, 2008] Grzegorczyk, M. and Husmeier, D. (2008). Improving the structure mcmc sampler for bayesian networks by introducing a new edge reversal move. *Machine Learning*, 71:265–305.

[Hall and Levison, 1990] Hall, P. A. and Levison, D. A. (1990). Review: assessment of cell proliferation in histopathological material. *Journal of Clinical Pathology*, 43:184–192.

[Hamilton and Aaltonen, 2000] Hamilton, S. R. and Aaltonen, L. A. (2000). *Pathology and genetics of tumours of the digestive system*, pages 124–165. IARC Press.

[Hanahan and Weinberg, 2000] Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell Review.*, 100:57–70.

[Hanahan and Weinberg, 2011] Hanahan, D. and Weinberg, R. A. (2011). The hallmarks of cancer. *Cell Review.*, 144:646–674.

[Hand et al., 2001] Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*. The MIT Press", first edition.

[Harrell, 2001] Harrell, F. E. J. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*, pages 56–87. Springer, first edition.

[Harrell, 2012a] Harrell, F. E. J. (2012a). Package rms. *CRAN*, pages 5–24.

[Harrell, 2012b] Harrell, F. E. J. (2012b). *Regression modeling strategies*, pages 34–98.

[Harrell et al., 1982] Harrell, F. E. J., Califf, R. M., Pryor, D. B., and Rosati, K. L. L. R. A. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association*, 247:2543–2546.

[Harrell. et al., 1996] Harrell., F. E. J., Lee, K. L., and Mark, D. B. (1996). Tutorial in biostatistics: Multivariate prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15:361–387.

[Hartemink, 2005] Hartemink, A. J. (2005). Reverse engineering gene regulatory networks. *Nature Biotechnology*, 23:554–555.

[Hartemink et al., 2001] Hartemink, A. J., Gifford, D. K., Jaakkola, T. S., and Young, R. A. (2001). Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pacific Symposium on Biocomputing*, pages 422–433.

[Hartemink et al., 2002] Hartemink, A. J., Gifford, D. K., Jaakkola, T. S., and Young, R. A. (2002). Combining location and expression data for principled discovery of genetic regulatory network models. *Pacific Symposium on Biocomputing*, pages 437–449.

[Hartwell and Kastan, 1994] Hartwell, L. and Kastan, M. (1994). Cell cycle control and cancer. *Science*, 266(5192):1821–1828.

[Hausser, 2006] Hausser, J. (2006). *Improving Entropy Estimation and the Inference of Genetic Regulatory Networks*. PhD thesis, Institut National des Science Appliquées; 28-65, Lyon.

[Heckerman, 1996] Heckerman, D. (1996). A tutorial on learning with bayesian networks. Technical Report MSR-TR-95-06, Microsoft research.

[Heckerman and Geiger, 1995] Heckerman, D. and Geiger, D. (1995). Learning bayesian networks: A unification for discrete and gaussian domain. *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 274–284.

[Heckerman et al., 1995] Heckerman, D., Geiger, D., and Chickering, D. M. (1995). Learning bayesian networks: The combination of knowledge and statistical data. Technical Report MSR-TR-94-09, Microsoft research.

[Henley and Kumamoto, 1999] Henley, E. J. and Kumamoto, H. (1999). *Probabilistic risk assessment: reliability engineering, design, and analysis*, pages 67–124. IEEE Press, third edition.

[Hippert et al., 2006] Hippert, M. M., S.O'Toole, P., and Thorburn, A. (2006). Autophagy in cancer: good, bad, or both? *Cancer Research*, 66(19):9349–9351.

[Hitchins, 2007] Hitchins, D. K. (2007). *Systems Engineering: A 21st Century Systems Methodology*. John Wiley.

[Hoeting et al., 1999] Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). The role of occam's razor in knowledge discovery. *Statistical Sciences*, 14:382–417.

[Holland, 2008] Holland, S. M. (2008). *Principal Component Analysis (PCA)*, pages 2–32.

[Hong et al., 2010] Hong, W. K., Jr., R. C. B., Hait, W. N., Kufe, D. W., Pollock, R. E., Weichselbaum, R. R., Holland, J. F., and III, E. F. (2010). *Holland-Frei Cancer Medicine*. American Association of Cancer Research (AACR), eighth edition.

[Hunter, 2009] Hunter, L. E. (2009). *The Processes of Life*, pages 34–56. MIT Press, first edition.

[Hynes and MacDonald, 2009] Hynes, N. E. and MacDonald, G. (2009). Erbb receptors and signaling pathways in cancer. *Cell Biology*, 21:177–184.

[Ilina and Friedl, 2009] Ilina, O. and Friedl, P. (2009). Mechanisms of collective cell migration at a glance. *Journal of Cell Science*, 122:3203–3208.

[Janez Demsar and Curk, 2004] Janez Demsar, Blaz Zupan, G. L. and Curk, T. (2004). Orange: From experimental machine learning to interactive data mining. *KPDD 2004*, 3202:537–539.

[Janzi et al., 2005] Janzi, M., Ödling, J., Pan-Hammarström, Q., Sundberg, M., Lundeberg, J., Uhlen, M., Hammarström, L., and Nilsson, P. (2005). Serum microarrays for large scale screening of protein levels. *Molecular and Cellular Proteomics*, 4:1942–1947.

[Jebara, 2002] Jebara, T. (2002). *Discriminative, Generative and Imitative Learning*. PhD thesis, Massachusetts Institute of Technology (MIT), US; 34-54.

[Jemal et al., 2011] Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., and Forman, D. (2011). Global cancer statistics. *CA - A Cancer Journal for Clinicians*, 61:69–90.

[Kalbfleisch and Prentice, 2002] Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, pages 78–132. Wiley, second edition.

[Kallioniemi et al., 2001] Kallioniemi, O.-P., Wagner, U., Kononen, J., and Sauter, G. (2001). Tissue microarray technology for high-throughput molecular profiling of cancer. *Human Molecular Genetics*, 10:657–662.

[Kalluri, 2003] Kalluri, R. (2003). Angiogenesis: Basement membranes: structure, assembly and role in tumour angiogenesis. *Nature Reviews Cancer*, 3:422–433.

[Kalluri and Weinberg, 2009] Kalluri, R. and Weinberg, R. A. (2009). The basics of epithelial-mesenchymal transition. *The Journal of Clinical Investigation*, 119(6):1420–1428.

[Kalluri and Zeisberg, 2006] Kalluri, R. and Zeisberg, M. (2006). Fibroblasts in cancer. *Nature Reviews Cancer*, 6:392–401.

[Katz et al., 2010] Katz, E., Marshall, S. D., Sims, A., Faratian, D., Li, J., Smith, E., Quinn, J., Edward, M., Meehan, R., Evans, E., Langdon, S., and Harrison, D. (2010). A gene on the her2 amplicon, c35, is an oncogene in breast cancer whose actions are prevented by inhibition of syk. *British Journal of Cancer*, 103:401–410.

[Katz et al., 2011] Katz, E., Verleyen, W., Blackmore, C. G., Edward, M., Smith, V. A., and Harrison, D. J. (2011). An analytical approach differentiates between individual and collective cancer invasion. *Analytical Cellular Pathology (ACP)*, 34:35–48.

[Kendall, 2004] Kendall, M. (2004). *Multivariate Analysis*, pages 23–43. Charles Griffin, fourth edition.

[Kleinbaum and Klein, 2005] Kleinbaum, D. G. and Klein, M. (2005). *Survival analysis: A Self-Learning Text*, pages 278–351. Springer, second edition.

[Koller and Friedman, 2009a] Koller, D. and Friedman, N. (2009a). *Bayesian Data Analysis*, pages 43–102. MIT press, first edition.

[Koller and Friedman, 2009b] Koller, D. and Friedman, N. (2009b). *Probabilistic Graphical Models: Principles and Techniques*, pages 145–231. MIT press, first edition.

[Kondo et al., 2005] Kondo, Y., Kanzawa, T., Sawaya, R., and Kondo, J. (2005). The role of autophagy in cancer development and response to therapy. *Nature: Reviews Cancer*, 5:726–734.

[Kononen et al., 1998] Kononen, J., Bubendorf, L., Kallioniemi, A., M, B., P., S., S., L., J., T., M.J., M., G., S., and O.P., K. (1998). Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nature Medicine*, 4:844–847.

[Korb and Nicholson, 2004] Korb, K. B. and Nicholson, A. E. (2004). *Bayesian Artificial Intelligence*, pages 113–145. Chapman and Hall, second edition.

[Kudoh, 2000] Kudoh, T. (2000). Tinysvm : Support vector machines. *http://cl.aist-nara.ac.jp/ taku-ku/software/TinySVM/index.html*.

[Kullari, 2003] Kullari, R. (2003). Basement membranes: structure, assembly and role in tumour angiogenesis. *Nature Reviews: Cancer*, 3:422–433.

[Lahrmann et al., 2011] Lahrmann, B., Halama, N., Sinn, H.-P., Schirmacher, P., Jaeger, D., and Grabe, N. (2011). Automatic tumor-stroma separation in fluorescence tmas enables the quantitative high-throughput analysis of multiple cancer biomarkers. *PLoS One*, 6:e28048.

[Langdon and Smyth, 2008] Langdon, S. P. and Smyth, J. F. (2008). Hormone therapy for epithelial ovarian cancer. *Current Opinion in Oncology*, 20:548–553.

[Larson, 1931] Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. *Journal of Educational Psychology*, 22(1):45–55.

[Li and Gui, 2004] Li, H. and Gui, J. (2004). Partial cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics*, 20:208–215.

[Lu et al., 2004] Lu, K. H., Patterson, A. P., Wang, L., Marquez, R. T., Atkinson, E. N., Baggerly, K. A., Ramoth, L. R., Rosen, D. G., Liu, J., Hellstrom, I., Smith, D., Hartmann, L., Fishman, D., Berchuck, A., Schmandt, R., Whitaker, R., Gershenson, D. M., Mills, G. B., and Robert C. Bast, J. (2004). Selection of potential markers for epithelial ovarian cancer with gene expression arrays and recursive descent partition analysis. *Clinical Cancer Research*, 10:3291–3300.

[Ludwig and Weinstein, 2005] Ludwig, J. A. and Weinstein, J. N. (2005). Biomarkers in cancer staging, prognosis, and treatment selection. *Nature Reviews Cancer*, 5:845–856.

[M. A. Aizerman and Rozonoér, 1964] M. A. Aizerman, E. M. B. and Rozonoér, L. I. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837.

[MacKay, 2004] MacKay, D. J. C. (2004). *Information Theory, Inderence, and Learning Algorithms*, pages 78–105. University Press Cambridge, first edition.

[Madigan and Raftery, 1994] Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of American Statistical Association*, 89:1535–1546.

[Mankoo et al., 2011] Mankoo, P. K., Shen, R., Schultz, N., Levine, D. A., and Sander, C. (2011). Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. *PLoS ONE*, 6.

[Mathew et al., 2007] Mathew, R., Karantza-Wadsworth, V., and White, E. (2007). Role of autophagy in cancer. *Nature Reviews: Cancer*, 7:961–967.

[McCafferty et al., 2009] McCafferty, M. P., Healy, N. A., and Kerin, M. J. (2009). Breast cancer subtypes and molecular biomarkers. *Diagnistic Histopathology*, 15:485–489.

[Meleth et al., 2007] Meleth, S., Chatla, C., Katkoori, V. R., Anderson, B., Hardin, J. M., Jhala, N. C., Bartolucci, A., Grizzle, W. E., and Manne, U. (2007). Comparison of predicted probabilities of proportional hazards regression and linear discriminant analysis methods using a colorectal cancer molecular biomarker database. *Cancer Informatics*, 3:115–221.

[Metz, 1978] Metz, C. E. (1978). Basic principles of roc analysis. *Seminars in Nuclear Medicine*, 8:283–298.

[Meuller and Fusenig, 2004] Meuller, M. M. and Fusenig, N. E. (2004). Friends or foes – bipolar effects of the tumour stroma in cancer. *Nature Reviews: Cancer*, 4:839–849.

[Meyer et al., 2012] Meyer, D., Dimitriadou, E., Hornik, K., Leisch, F., and Weingessel, A. (2012). *Package: e1071*. R package version 1.6.

[Meyer et al., 2008] Meyer, P. E., Lafitte, F., and Bontempi, G. (2008). minet: A r/bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, 9:462:1134–1145.

[Mitchell, 1997] Mitchell, T. M. (1997). *Machine Learning*, pages 11–217. McGraw-Hill, first edition.

[Moreau and Tranchevent, 2012] Moreau, Y. and Tranchevent, L.-C. (2012). Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature Reviews Genetics*, 13:523–536.

[Murphy et al., 2012] Murphy, S. L., Xu, J., and Kochanek, K. D. (2012). National vital statistics report. 60:67–69.

[Neal, 1995] Neal, R. M. (1995). *Bayesian Learning for Neural Networks*. PhD thesis, University of Toronto, Toronto.

[Nevins, 2007] Nevins, J. R. (2007). New breast cancer genes-discovery at the intersection of complex data sets. *Cancer Cell*, 12:497–499.

[Noble, 1961] Noble, D. (1961). A modification of the hodgkin-huxley equations applicable to purkinje fibre action and pace-maker potentials. *The Journal of Physiology*, 160:317–352.

[Noble, 2006] Noble, D. (2006). *The music of life: Biology Beyond the Genome*, pages 23–67. Oxford Univerity Press, Oxford UK, first edition.

[Nodelman et al., 2002] Nodelman, U., Shelton, C. R., and Koller, D. (2002). Learning continuous time bayesian networks. *Proceedings of the Nineteenth conference on Uncertainty in artificial intelligence*, pages 451–458.

[Olsson, 2002] Olsson, U. (2002). *Generalized Linear Models: An Applied Approach*, pages 12–96. Lund.

[Palsson, 2006] Palsson, B. (2006). *Systems Biology: Properties of reconstructed networks*, pages 14–118. Cambridge University Press, first edition.

[Park, 2007] Park, M. Y. (2007). l_1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society*, 69:659–677.

[Pearl, 1988] Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*, pages 124–134. Morgan Kaufmann.

[Press et al., 2007] Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (2007). *Numerical Recipes*. Cambridge University Press, third edition.

[Raftery, 1995] Raftery, A. E. (1995). Bayesian model selection in social research. *Socialogical Methodology*, 25:111–163.

[Rakha et al., 2008a] Rakha, E., Putti, T., El-Rehim, D. A., Paish, C., Green, A., Powe, D., Lee, A., Robertson, J., and Ellis, I. (2008a). Morphological and immunophenotypic analysis of breast carcinomas with basal and myoepithelial differentiation. *Journal of Pathology*, 208:2568–2581.

[Rakha et al., 2008b] Rakha, E. A., Reis-Filho, J. S., and Ellis, I. O. (2008b). Basal-like breast cancer: A critical review. *Journal of Clinical Oncology*, 26:2568–2581.

[Ransohoff, 2004] Ransohoff, D. F. (2004). Rules of evidence for cancer molecular-marker discovery and validation. *Nature Reviews Cancer*, 4:309–314.

[Ransohoff, 2005] Ransohoff, D. F. (2005). Bias as a threat to the validity of cancer molecular-marker research. *Nature Reviews Cancer*, 5:142–149.

[Reis-Filho, 2009] Reis-Filho, J. S. (2009). Next-generation sequencing. *Beast Cancer Research*, 11.

[Richardson and Domingos, 2006] Richardson, M. and Domingos, P. (2006). Markov logic networks. *Machine Learning*, 62:107–136.

[Roberts et al., 2012] Roberts, N. J., Vogelstein, J. T., Parmigiani, G., Kinzler, K. W., Vogelstein, B., and Velculescu, V. E. (2012). The predictive capacity of personal genome sequencing. *Science Translational Medicine*, pages 1–12.

[Sachs et al., 2005] Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308:523–529.

[Sanchez-García, 2009] Sanchez-García, I. (2009). The crossroads of oncogenesis and metastasis. *The New England Journal of Medicine*, 360:297–299.

[Schena et al., 1995] Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270:467–470.

[Schlabach et al., 2008] Schlabach, M. R., Luo, J., Solimini, N. L., Hu, G., Xu, Q., Li, M. Z., Zhao, Z., Smogorzewska, A., Sowa, M. E., Ang, X. L., Westbrook, T. F., Liang, A. C., Chang, K., Hackett, J. A., Harper, J. W., Hannon, G. J., and Elledge, S. J. (2008). Cancer proliferation gene discovery through functional genomics. *Science*, 319:620–624.

[Scholköpf, 2001] Scholköpf, B. (2001). The kernel trick for distances. *Advances in Neural Information Processing Systems*, 13:123–131.

[Scholköpf et al., 2000] Scholköpf, B., Williamson, R. C., and Bartlett, P. L. (2000). New support vector algorithms. *Neural Computation*, 12:1207–1245.

[Schoumacher et al., 2010] Schoumacher, M., Goldman, R. D., Louvard, D., and Vignjevic, D. M. (2010). Actin, microtubules, and vimentin intermediate filaments cooperate for elongation of invadopodia. *The Journal of Cell Biology*, 189:541–556.

[Schuster, 2008] Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nature Methods*, 5:16–18.

[Schwartz, 1978] Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

[Scutari, 2010] Scutari, M. (2010). Learning bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):1–22.

[Sengupta, 2006] Sengupta, A. (2006). *Choas, Nonlinearity, complexity: The Dynamic Paradigm of Nature*. Springer.

[Serlin and Levin, 1985] Serlin, R. C. and Levin, J. R. (1985). Teaching how to derive directly interpretable coding schemes for multiple regression analysis. *Journal of Educational Statistics*, 10:223–238.

[Sethi and Kang, 2011] Sethi, N. and Kang, Y. (2011). Unraveling the complexity of metastasis – molecular understanding and targeted therapies. *Nature: Reviews Cancer*, 11:735–748.

[Silander et al., 2007] Silander, T., Kontkanen, P., and Myllymaki, P. (2007). On sensitivity of the map bayesian network structure to the equivalent sample size parameter. pages 145–163.

[Sing et al., 2005] Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). Rocr: visualizing classifier performance in r. *Bioinformatics*, 21:3940–3941.

[Smith, 2002] Smith, L. I. (2002). *A tutorial on Principle Components Analysis*, pages 3–13.

[Smola and Schölkopf, 2004] Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14:199–222.

[Solomonoff, 1956] Solomonoff, R. J. (1956). An inductive inference machine. Technical report, Technical Research Group, New York City.

[Sorlie et al., 2006] Sorlie, T., Perou, C. M., Fan, C., Geisler, S., Aas, T., Nobel, A., Anker, G., Akslen, L. A., Botstein, D., Borresen-Dale, A.-L., , and Lonning, P. E. (2006). Gene expression profiles do not consistently predict the clinical treatment response in locally advanced breast cancer. *Molecular Cancer Therapeutics*, 5:2914–2918.

[Sorlie et al., 2001] Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eise, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Lnning, P. E., and Brresen-Dale, A.-L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *PNAS*, 98:10869–10874.

[Spiegelhalter et al., 1993] Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L., and Cowell, R. G. (1993). Bayesian analysis in expert systems. *Journal of Statistical Science*, 8:219–283.

[Spurrier et al., 2008] Spurrier, B., Ramalingam, S., and Nishizuka, S. (2008). Reverse-phase protein lysate microarrays for cell signaling analysis. *Nature Protocols*, 3:1796–1808.

[Steck and Jaakkola, 2006] Steck, H. and Jaakkola, T. S. (2006). Predictive discretization during model selection.

[Steuer et al., 2002] Steuer, R., Kurths, J., Daub, C., Weise, J., and Selbig, J. (2002). The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics*, 18-2:S231–S240.

[Stevenson, 2009] Stevenson, M. (2009). Introduction to survival analysis. Technical report, EpiCentre, IVABS, Massey University.

[Sundström, 2010] Sundström, S. (2010). Coding in multiple regression analysis: A review of popular coding techniques. Technical report, Uppsala Universitet.

[Tibes et al., 2005] Tibes, R., Qui, Y., Hennessy, B., Andreeff, M., Mills, G. B., and Kornblau, S. M. (2005). Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. *Molecular Cancer Therapeutics*, 5:2512–2521.

[Tibshirani, 1995] Tibshirani, R. (1995). Regression shrinkage and selection via lasso. *Journal of the Royal Statistical Society*, 58:267–288.

[Tibshirani, 1997] Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16:385–395.

[Tikhonov, 1995] Tikhonov, N. (1995). *Numerical Methods for the Solution of Ill-Posed Problems*, pages 34–76. Springer, first edition.

[Timpson et al., 2011] Timpson, P., McGhee, E. J., Morton, J. P., von Kriegsheim, A., Schwarz, J. P., Karim, S. A., Doyle, B., Quinn, J. A., Carragher, N. O., Edward, M., Olson, M. F., Frame, M. C., Brunton, V. G., Sansom, O. J., and Anderson, K. I. (2011). Spatial regulation of rhoa activity during pancreatic cancer cell invasion driven by mutant p53. *Cancer Research*, 71:747–757.

[Tobias, 2002] Tobias, R. D. (2002). An introduction to partial least squares regression.

[Tothill et al., 2008] Tothill, R. W., Tinker, A. V., George, J., Brown, R., Fox, S. B., Lade, S., Johnson, D. S., Trivett, M. K., Etemadmoghadam, D., Locandro, B., Traficante, N., Fereday, S., Hung, J. A., Chiew, Y.-E., Haviv, I., Gertig, D., deFazio, A., and Bowtell, D. D. (2008). Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clinical Cancer Research*, 14:5198–5208.

[Tsamardinos et al., 2006] Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006). The max-min hill-climbing bayesian network structure learning. *Machine Learning*, 65:31–78.

[UKCR, 2012] UKCR (2012). Cancer research uk: common causes of mortality in uk (2007-2009). http://info.cancerresearchuk.org/cancerstats/mortality/.

[van den Berg, 2001] van den Berg, G. J. (2001). *Duration Models: Specification, Identification, and Multiple Durations*, pages 78–115. Elsevier, five edition.

[van Houwelingen J.C. and le Cessie S., 1990] van Houwelingen J.C. and le Cessie S. (1990). Predictive value of statistical models. *Statistics in Medicine*, 8:1303–1325.

[van't Veer et al., 2005] van't Veer, L. J., Paik, S., and Hayes, D. F. (2005). Gene expression profiling of breast cancer: A new tumor marker. *Journal of Clinical Oncology*, 23:1631–1635.

[Vapnik, 2010a] Vapnik, V. N. (2010a). *The Nature of Statistical Learning Theory*, pages 56–72. Springer, second edition.

[Vapnik, 2010b] Vapnik, V. N. (2010b). *Statistical Learning Theory*. Springer, second edition.

[Vazquez-Martin et al., 2009] Vazquez-Martin, A., Oliveras-Ferraros, C., and Menendez, J. A. (2009). Autophagy facilitates the development of breast cancer resistance to the anti-her2 monoclonal antibody trastuzumab. *Plos One*, 4(7).

[Verheul and Pinedo, 2007] Verheul, H. M. W. and Pinedo, H. M. (2007). Possible molecular mechanisms involved in the toxicity of angiogenesis inhibition. *Nature Reviews: Cancer*, 7:475–485.

[Wagner, 2010] Wagner, A. (2010). *Robustness and Evolvability in Living Systems*, pages 12–48. Princeton Press, first edition.

[Wang et al., 2010] Wang, X., He, L., Wu, Y. I., Hahn, K. M., and Montell, D. J. (2010). Light-mediated activation reveals a key role for rac in collective guidance of cell movement in vivo. *Nature: Cell Biology*, 12:591–598.

[Weaver, 2006] Weaver, A. M. (2006). Invadopodia: specialized cell structures for cancer invasion. *Clinical and Experimental Metastasis*, 23:97–105.

[Westerhoff and Palsson, 2004] Westerhoff, H. V. and Palsson, B. O. (2004). The evolution of molecular biology into systems biology. *Nature: Biotechnology*, 22:1249–1252.

[Yan et al., 2004] Yan, L., Verbel, D., and Saidi, O. (2004). Predicting prostate cancer recurrence via maximizing the concordence index. *International Conference on Knowledge Discovery and Data Mining (KDD)*.

[Yarden and Sliwkowski, 2001] Yarden, Y. and Sliwkowski, M. X. (2001). Untangling the erbb signalling network. *Nature Reviews Molecular Cell Biology*, 2:127–137.

[Yilmaz et al., 2007] Yilmaz, M., Christofori, G., and Lehembre, F. (2007). Distinct mechanisms of tumor invasion and metastatis. *Trends in Molecular Medicine*, 13:535–541.

[Yu, 2005] Yu, J. (2005). *Developing Bayesian Network inference algorithms to predict causal functional pathways in biological systems*. PhD thesis, Duke University; 102-109, US.

[Yu et al., 2004] Yu, J., Smith, V. A., Wang, P. P., Hartemink, A. J., and Jarvis, E. D. (2004). Advances to bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, 20(18):3594–3603.

[Yuri, 2010] Yuri, L. (2010). What are the hallmarks of cancer? *Nat Rev Cancer*, 10(4):232–233.

7
Appendix A: data sets

(Spreadsheet embargoed along with Chapters 3 & 5 at the author's request)

This appendix will provide an overview of the data sets that are used in course of this doctorate. All data set are collected in a spreadsheet: PhD_Wim_Verleyen.xls. All the data was collect at the laboratory of my collaborators at the Division of Pathology at the University of Edinburgh.

7.1 Edinburgh Ovarian Cancer Register (EOCR)

From the Edinburgh Ovarian Cancer Register (EOCR), there are two data sets related to the study explained in chapter 3 on page 89: (1) the original data set used for feature selection and training set for the 1YM-PFS and 3YM-OS classifiers and (2) additional validation data set for the validation of 1YM-PFS and 3YM-OS classifiers.

7.1.1 Original data set for feature selection and training set for 1YM-PFS and 3YM-OS classifiers

EOCR_original sheet contains the data with no missing values. This data contains the following clinicopathological variables:

Age age of the patient (minimum age = 30 and maximum age = 86).

AgeStratified age stratified by 50 years (1: patient is younger as 50 years and 2: patient is older as 50 years).

Stage the stage of the tumour (1: stage 1, 2: stage 2, 3: stage 3, and 4: stage 4).

Regimen the regimen prescription for a patient (1: platinum and 2: platinum combined with taxane).

HistologicalType histological type (1: papillary serous, 2: clear cell, 3: endometrioid, 4: mixed mullerian, 5: mucinous, and 6: adenocarcinoma).

7.1.2 Additional validation data set for the validation of 1YM-PFS and 3YM-OS classifiers

EOCR_validation sheet contains the additional data set (with no missing values) used for validation of the 1YM-PFS and 3YM-OS

Table 7.1: Frequency table of the histological types in each stage.

Histological type	Stage 1	Stage 2	Stage 3	Stage 4
Papillary serous	7	9	146	47
Clear cell	3	6	3	2
Endometrioid	5	7	36	10
Mixed mullerian	3	7	26	6
Mucinous	0	1	5	0
Adenocarcinoma	3	0	6	1

classifiers.

7.2 Tumour invasion data set

Tumour_invasion sheet contains the data used for the tumour invasion study in chapter 4 on page 127. This data set has no missing values.

8
Index